"十一五"国家重点图书

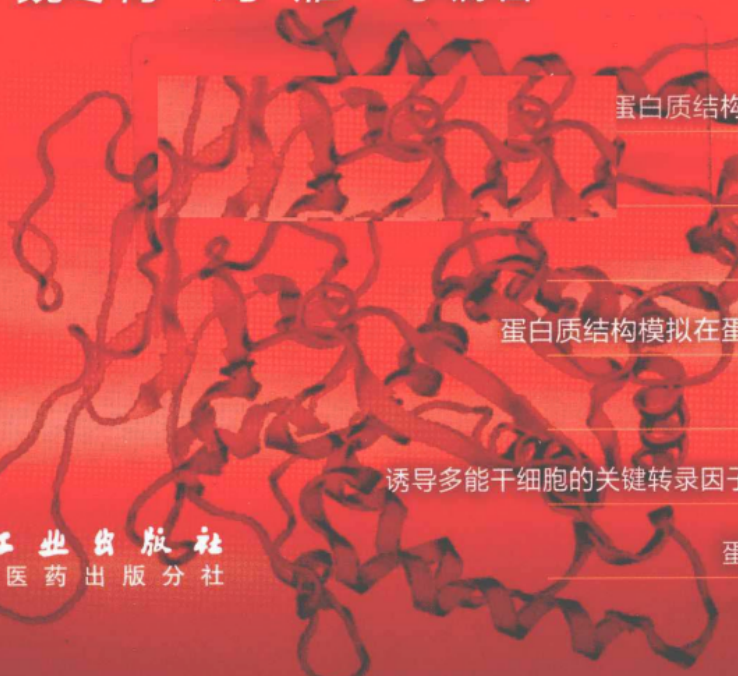
华夏英才基金学术文库

蛋 | 白 | 质 | 科 | 学 | 与 | 技 | 术 | 丛 | 书

# 蛋白质结构模拟与设计

## Protein Structure Modeling and Design

李荣秀　魏冬青　冯　雁　等编著



蛋白质结构研究的重大历史事件

蛋白质结构基础理论

从序列到结构

蛋白质结构模拟在蛋白研究中的应用实例

锌指核酸酶的构建

诱导多能干细胞的关键转录因子Oct4结构模拟与设计

蛋白质（酶）进化工程

化学工业出版社
生物·医药出版分社

蛋 | 白 | 质 | 科 | 学 | 与 | 技 | 术 | 丛 | 书

本书在扼要回顾蛋白质结构研究中的重要历史事件及概述蛋白质结构基础理论基础上，阐述蛋白质结构预测、蛋白质结构模拟、锌指核酸酶的构建、诱导多能干细胞的关键转录因子Oct4结构模拟与设计、蛋白质(酶)进化工程等内容。书中部分素材来源于作者自身研究成果和经验，内容较新。

作者撰写本书的目的之一在于引导从事蛋白质结构研究的科技工作者在解决具体问题借鉴前辈科学大师的有效逻辑、方法及强有力的意志。

ISBN 978-7-122-10183-9
9 787122 101839 >

销售分类建议：生物 / 生物技术
定价：49.00 元

"十一五"国家重点图书

华夏英才基金学术文库

蛋白质科学与技术丛书

# 蛋白质结构模拟与设计

## Protein Structure Modeling and Design

李荣秀　魏冬青　冯　雁　等编著

化学工业出版社
生物·医药出版分社
·北京·

**图书在版编目（CIP）数据**

蛋白质结构模拟与设计/李荣秀等编著. —北京：
化学工业出版社，2011.1
（“十一五”国家重点图书）
（华夏英才基金学术文库）
（蛋白质科学与技术丛书）
ISBN 978-7-122-10183-9

Ⅰ.蛋… Ⅱ.李… Ⅲ.蛋白质-生物结构
Ⅳ.Q510.1

中国版本图书馆 CIP 数据核字（2010）第 250908 号

责任编辑：傅四周　　文字编辑：焦欣渝
责任校对：边　涛　　装帧设计：关　飞

出版发行：化学工业出版社　生物·医药出版分社
（北京市东城区青年湖南街 13 号　邮政编码 100011）
印　　刷：北京永鑫印刷有限责任公司
装　　订：三河市万龙印装有限公司
710mm×1000mm　1/16　印张 10¾　彩插 3　字数 186 千字　2011 年 5 月北京第 1 版第 1 次印刷

购书咨询：010-64518888（传真：010-64519686）　售后服务：010-64518899
网　　址：http://www.cip.com.cn
凡购买本书，如有缺损质量问题，本社销售中心负责调换。

**定　　价：49.00 元**

**版权所有　违者必究**

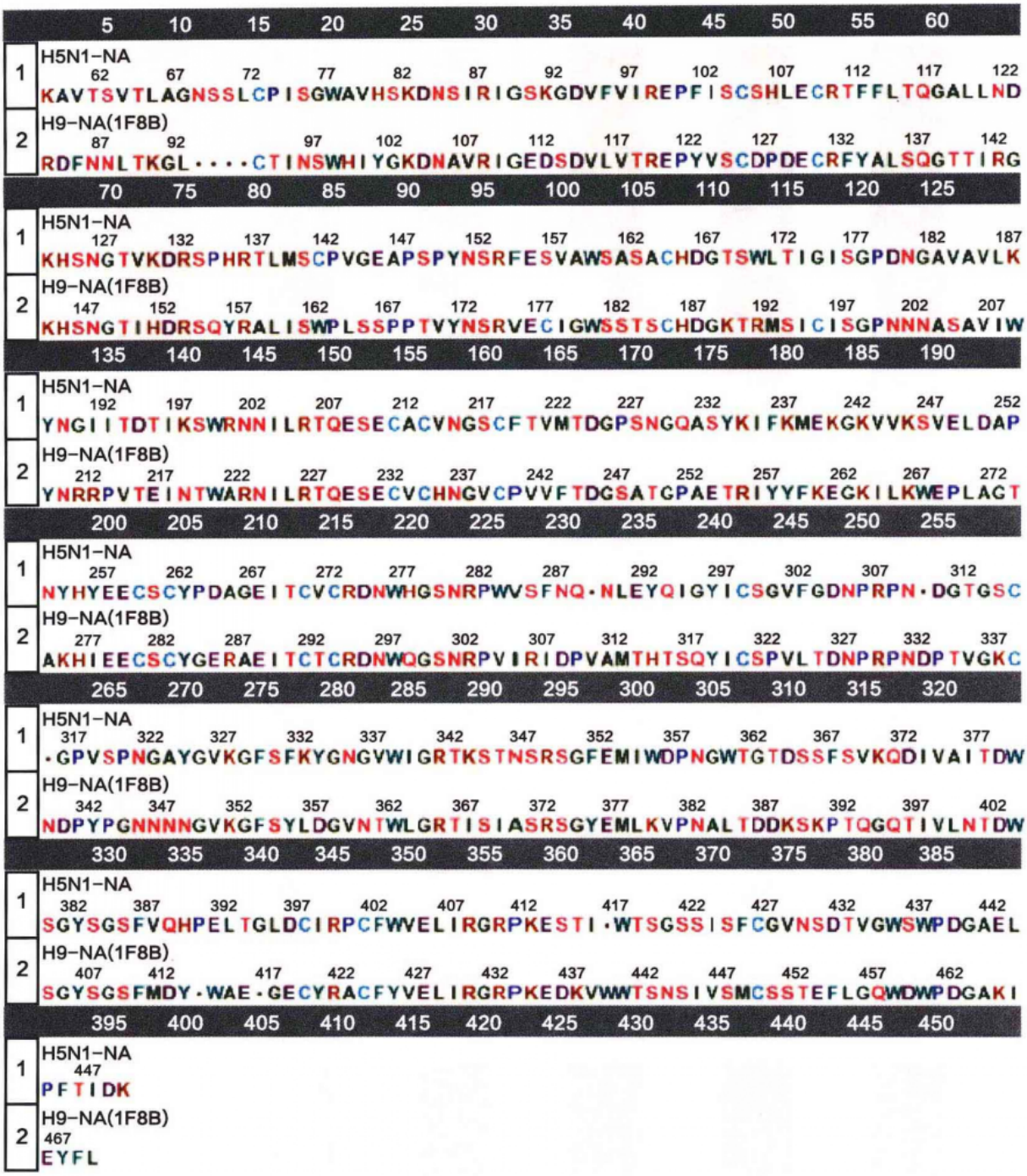



图 5.1　序列H5N1-NA（Accession No. Q5H89）和 N9-NA（Accession No. Q84070）比对

图中氨基酸颜色表示氨基酸的功能：酸性—红色；碱性—蓝色；中性亲水—粉红色；脂肪族—墨绿色；芳香族—青绿色；包含巯基—黄色；亚胺基—橙色

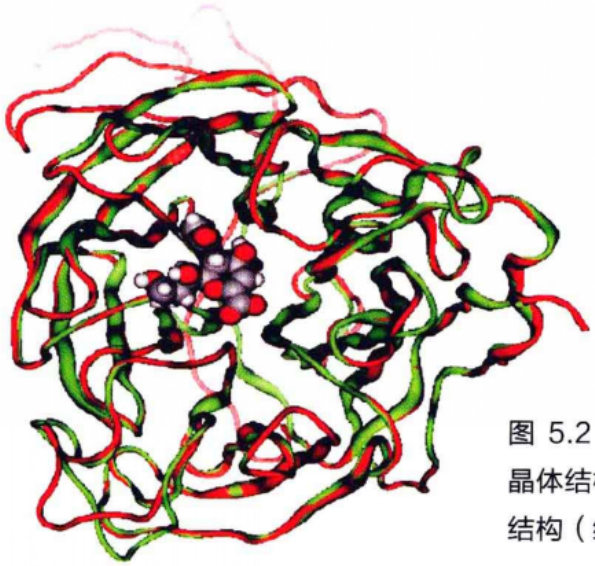
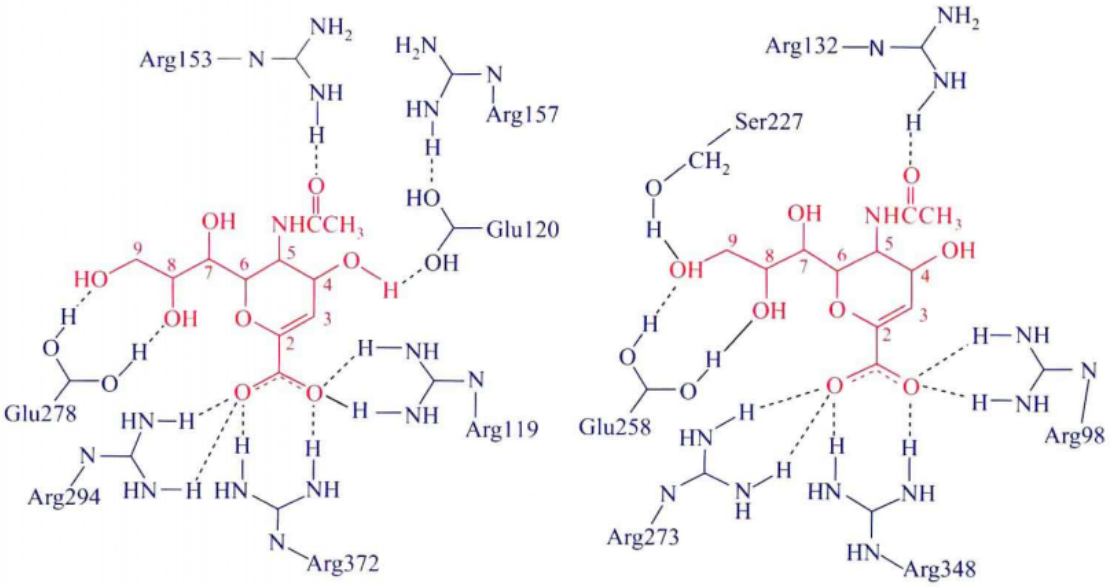


图 5.2　根据结合 DANA 的 N9-NA 复合物晶体结构（红色）构建的 H5N1-NA 的三维结构（绿色）




(a) N9-NA:DANA　　(b) H5N1-NA:DANA

图 5.4　配体 DANA（红色）分别与受体 N9-NA（a）和H5N1-NA（b）活性位点残基（蓝色）相互作用示意图

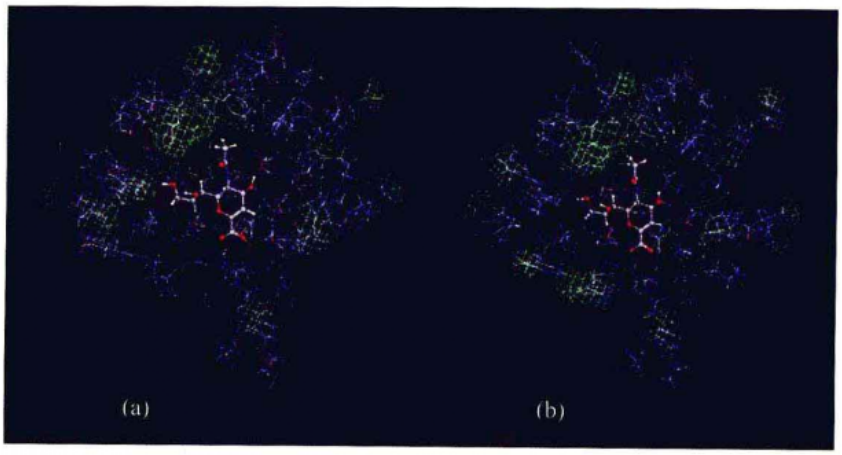



图 5.5　N9-NA（a）和 H5N1-NA（b）活性空腔的亲脂和亲水表面示意图

其中绿色表示亲脂性，蓝色表示亲水性

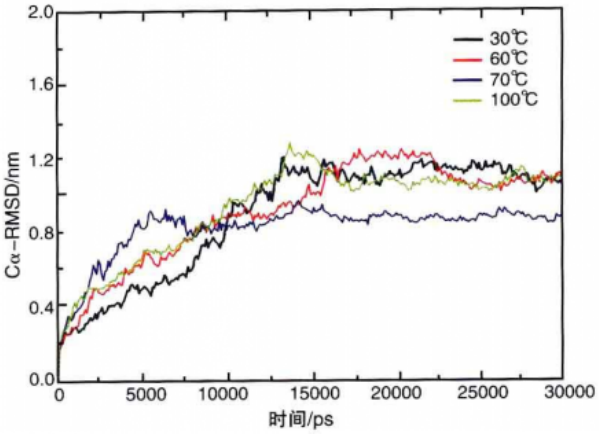



图 5.7　模拟过程中骨架碳原子的均方根偏差随时间变化示意图

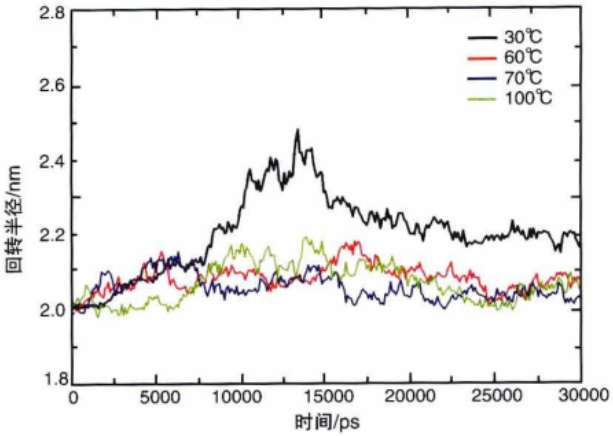



图 5.8　骨架碳原子的回转半径示意图

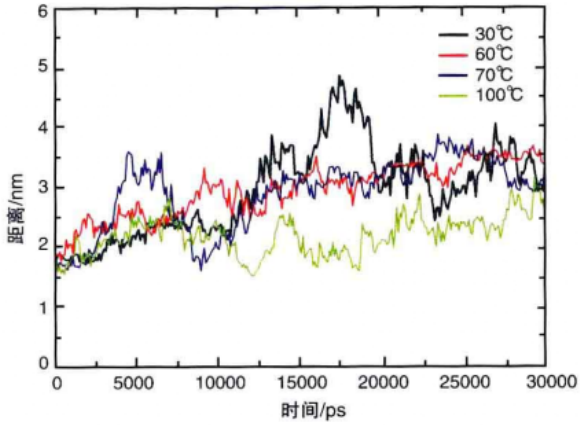



图 5.9　模拟过程中螺旋 6 与螺旋 9 之间的距离

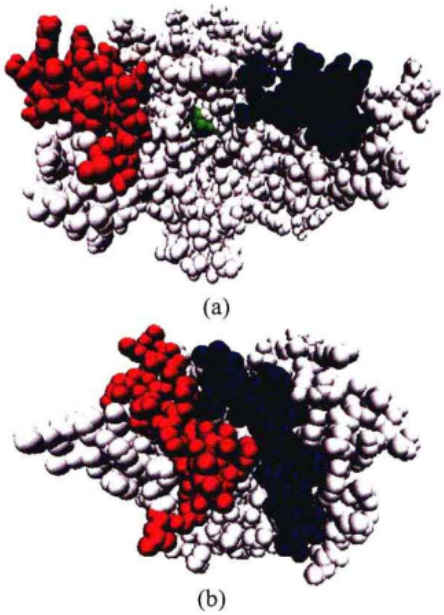



图 5.10　MD 模拟中 T1 脂肪酶打开（a）与关闭（b）的构象

红色代表螺旋9，蓝色代表螺旋6，绿色代表活性中心

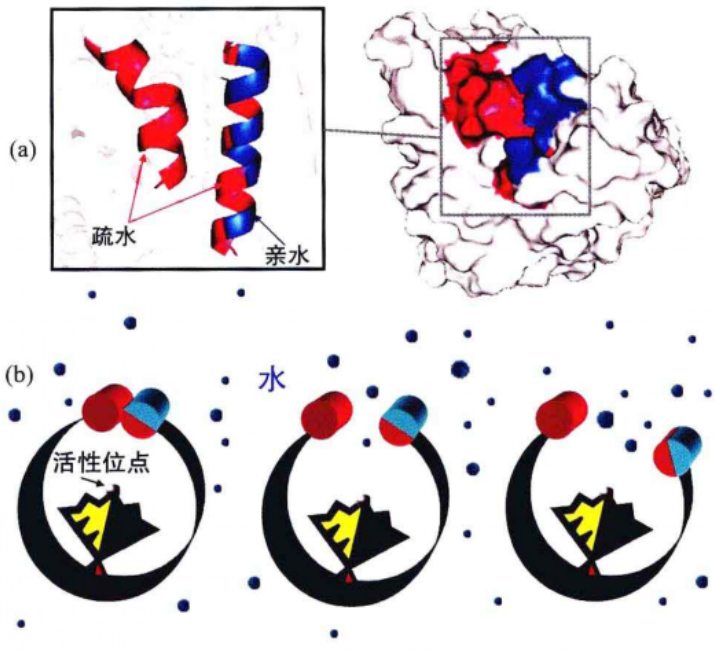



图 5.12　T1 脂肪酶结构变化的分子机理示意图

红色代表疏水，蓝色代表亲水

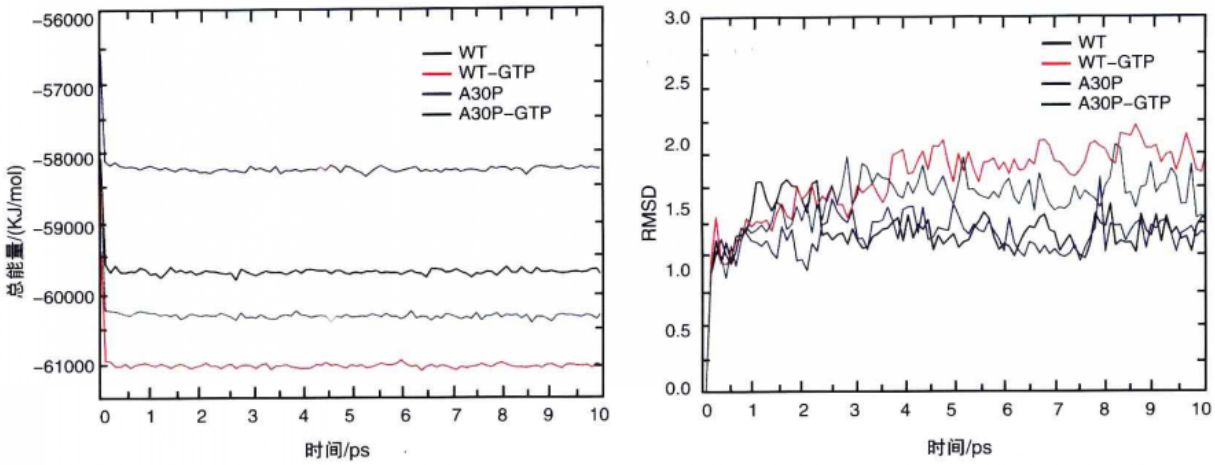



图 5.14　模拟过程中体系总能量和骨架碳原子的均方根偏差随时间变化示意图

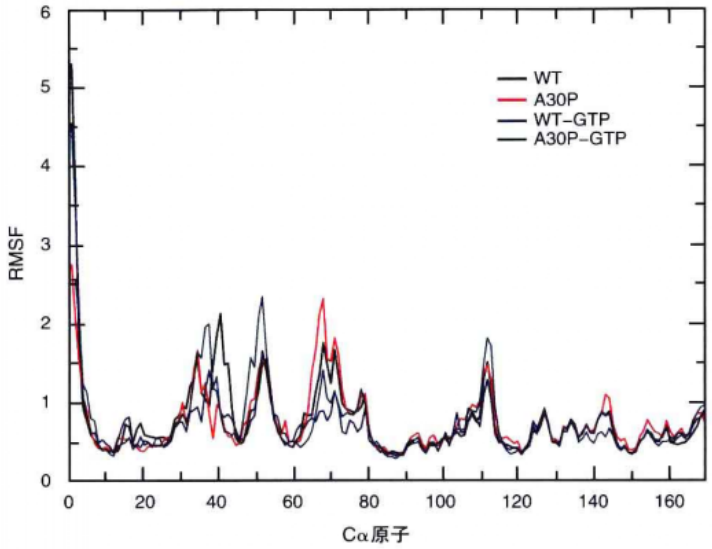



图 5.15　分子动力学过程中骨架碳原子的均方根波动示意图

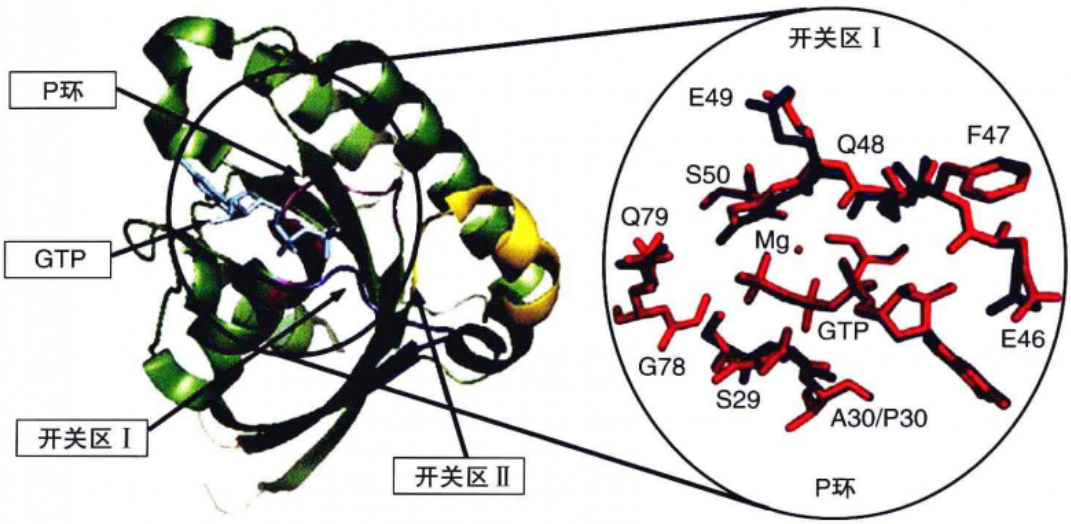



图 5.16　A30P突变所带来的结构变化示意图

图中野生型结构用蓝色棍状模型表示，而突变型结构则是用红色棍状模型表示

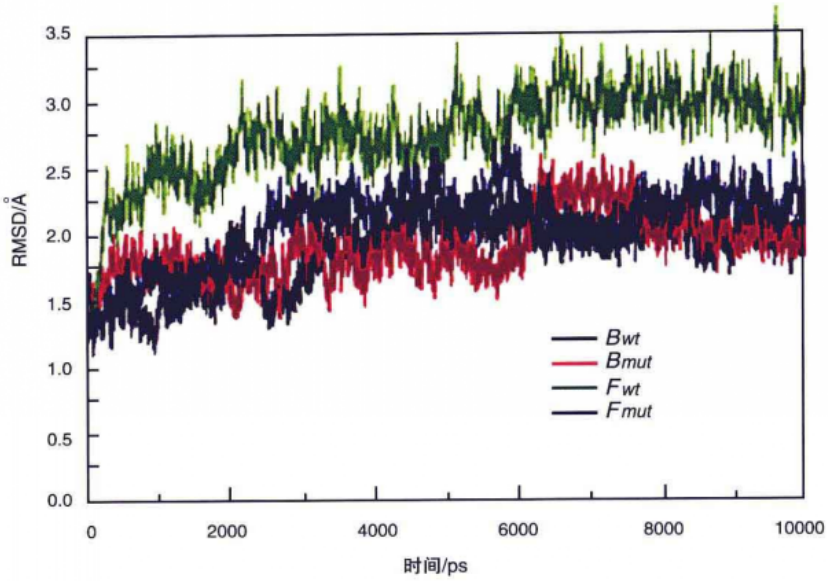



图 5.19　各体系蛋白酶骨架原子的 RMSD 随时间变化规律

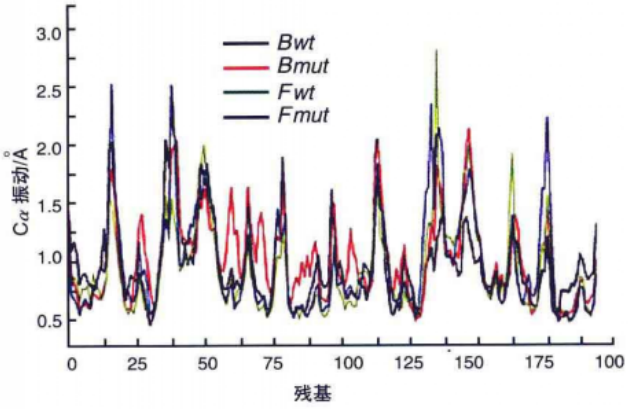



图 5.20　各体系中骨架碳原子在分子动力学模拟中的振动情况

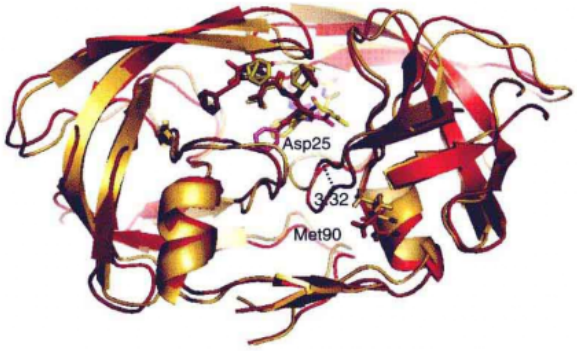



图 5.22　*Bmut* 蛋白酶 (红色) 与 *Bwt* (黄色) 蛋白酶的空间结构叠合

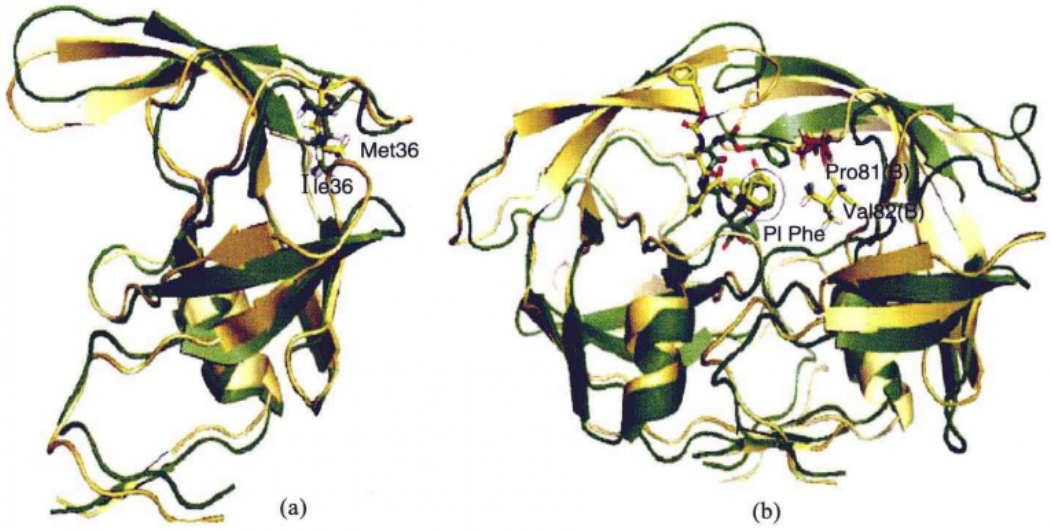



图 5.24　*Fwt* (绿色) 和 *Bwt* (黄色) 蛋白酶B链 (a) 和 S1/S1′ 疏水口袋 (b) 的空间结构叠合

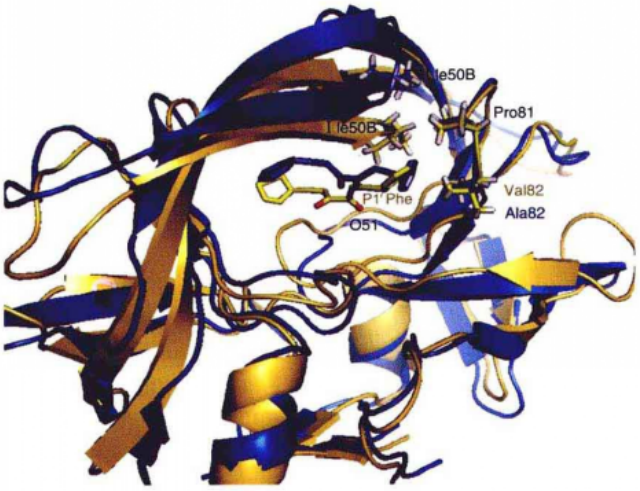



图 5.26　*Fmut* (蓝色) 和 *Bwt* (黄色)蛋白酶空间结构叠合

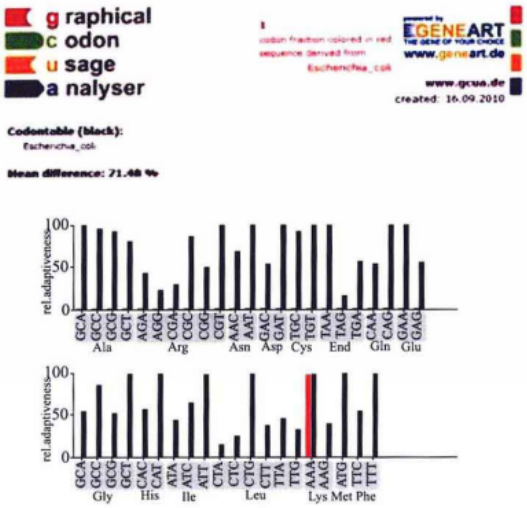



图 6.18　各种密码子在大肠杆菌中的使用频率

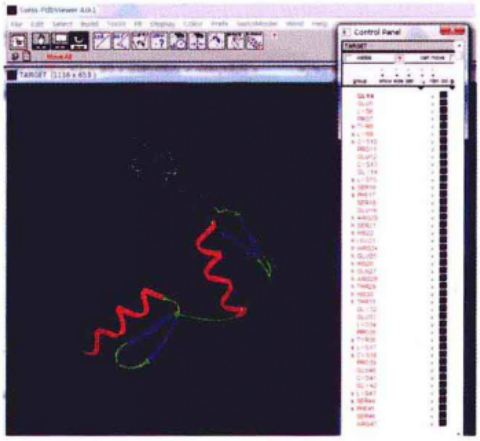


图 6.26　Deep View 软件工作界面

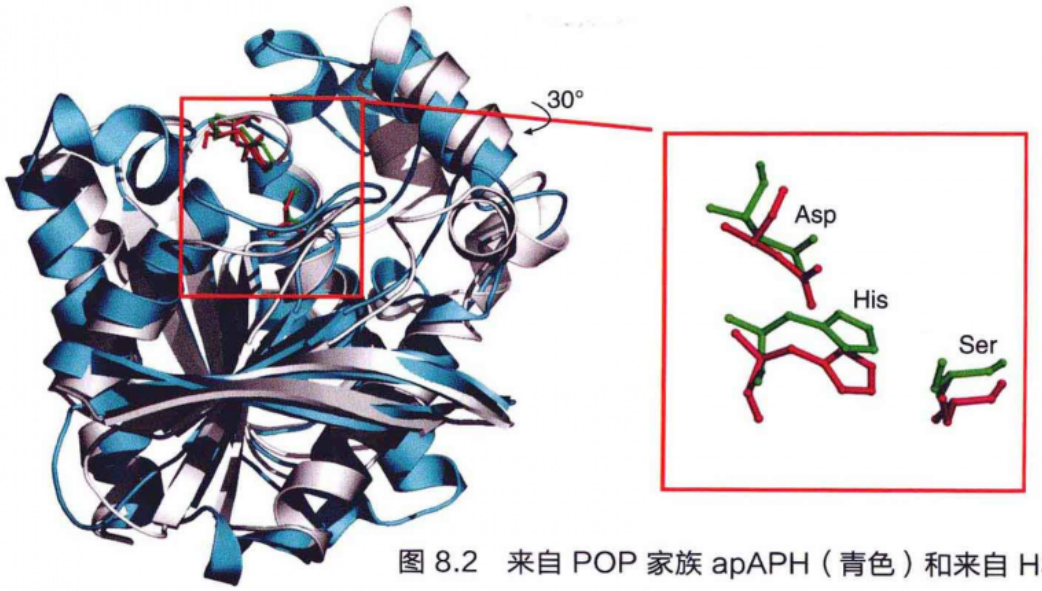



图 8.2　来自 POP 家族 apAPH（青色）和来自 HSL 家族的 AFEST（灰色）催化结构域和催化三联体的结构重叠

# 编 写 人 员

（按姓名汉语拼音排序）

陈仕高　程　超　冯　雁　李荣秀　连　鹏

王靖方　王　莹　魏冬青　仲从浩

# 丛书序

生命科学在21世纪已从基因组的研究进入功能基因组或结构基因组的研究，也就是说要在蛋白质结构的基础上研究基因编码的各种蛋白质的功能，并从传统的单个蛋白质的研究走向细胞内蛋白质群体的研究，从而更加深入地揭示生命活动的奥秘。这方面的工作构成了当今生命科学研究的前沿，即蛋白质科学。

人类对蛋白质重要性的认识源远流长。1878年恩格斯在《反杜林论》中就指出“生命是蛋白体的存在形式”。不过当时对蛋白质的本质和结构还知之甚少。不久，瑞典的Theodor Svedberg利用他首创的超离心技术才知道蛋白质分子是均一的并具有固定的分子量。此后，剑桥大学的Frederick Sanger利用纸电泳和纸层析测出了第一个蛋白质——胰岛素的一级结构；Max Perutz和John Kendrew利用他们首创的重原子同晶置换技术测出了第一个蛋白质——肌红蛋白晶体的三维结构。我国对蛋白质的研究起步也不晚。20世纪二三十年代，北京协和医学院吴宪领导的实验室对蛋白质的变性作用进行了深入系统的研究，提出蛋白质变性的实质是蛋白质分子从紧密而有序的结构变为散漫而无序的结构。吴宪的蛋白质变性学说对当前蛋白质分子折叠的研究仍然具有现实意义。及至20世纪五六十年代，中国科学院生物化学研究所、有机化学研究所和北京大学又通过社会主义大协作合成了胰岛素，而且得到了晶体。这不仅是世界上第一次在实验室合成了具有天然构象的蛋白质，而且彻底证明了“一级结构决定高级结构”这一重要原理。随后，我国又独立测出了胰岛素晶体的三维结构。正是由于国内外这些开创性的工作，才有了当前蛋白质科学的蓬勃发展。

回顾蛋白质科学发展的历程，我们可以看到先进技术对科学发展的无比重要性。除了上述的超离心、色谱、电泳、重原子同晶置换等技术以外，后来涌现的高压液相层析、二维凝胶电泳、毛细管电泳、质谱、生物芯片等技术又引领了蛋白质组学的研究，使我们有可能对细胞内的大量蛋白质群体进行综合研究。由此看来，科学与技术是相辅相成的，相互促进的，二者缺一不可。

在人类基因组已被基本破译的基础上，成千上万的蛋白质的结构与功能及其相互作用亟待阐明。蛋白质科学的研究成果将有助于阐明生命现象的本质和活动规律，为多种疾病的致病机理和防治提供理论依据。因此，蛋白质科学是发达国家激烈争夺的制高点，成为近年来生命科学不断取得重大突破的热点领域。我国

对蛋白质科学也给予了高度重视，《国家中长期科学和技术发展规划纲要(2006—2020年)》将“蛋白质研究”列为四项重大科学研究计划之一。在广大科研人员的共同努力下，我国在蛋白质研究领域已开始出现与国际先进水平基本同步的良好发展态势，取得了一些重要的理论和应用成果，陆续在“Nature”、“Science”、“Cell”等重要期刊上发表了一批有影响的论文。

正是在这样的背景下，化学工业出版社及时出版的这套《蛋白质科学与技术丛书》很有现实意义。丛书由相互独立而又彼此联系的各分册组成，主题包括“理论”和“技术”两个层面，现阶段以“技术”为主。归纳起来，这套丛书具有以下一些特色。

1. 选题涉及的范围较为广泛。从蛋白质的基础理论，到研究的各种技术方法，以至于前沿的蛋白质组研究，在丛书中都有体现，以满足不同方向、不同层次的科研和教学需要。

2. 编写队伍具有较高水准，由中国科学院、军事医学科学院、北京大学、南京大学等单位活跃在第一线的学术骨干为主编写。

3. 编写人员分布在国内外各研究机构，因而能扬各家之长，并体现国际化的学术合作。

值得欣慰的是，目前这套丛书已通过立项成为“十一五”国家重点图书出版项目，其编写计划也得到了学术界的大力支持，进展顺利。计划两年内陆续出版的第一批书目包括：

**蛋白质物理学概论**　　**蛋白质组学原理**
**蛋白质电泳技术指南**　　**蛋白质色谱分离技术**
**蛋白质质谱技术**　　**蛋白质微阵列**
**蛋白质高效表达技术**　　**蛋白质结构模拟与设计**

蛋白质科学的研究方兴未艾，今后的发展任重而道远。面对这一挑战，《蛋白质科学与技术丛书》的出版必将促进我国的蛋白质科学在已有的基础上进一步发扬光大。同时，随着这一领域的飞速发展，也希望这套丛书能不断推出后续的新篇和新人新作。

中国科学院院士
张友尚
2007年4月26日

# 前　言

蛋白质结构是生物科学的一个重要部分，与人类自身的重要哲学命题“生命是什么”密切相关。回顾这个研究领域的科学问题、研究方法、相应的理论和概念的发展以及相互促进的关系，能够启发对科学研究充满兴趣的青年及早考虑诸如此类的重要事项：什么是科学问题？科学问题是从哪里来的？重大科学发现的逻辑和巧妙是什么？科学问题的解决首先需要大胆的假设和推断，再借鉴和发展新方法和新工具，小心求证理论假设。

蛋白质结构研究的很多突破性发展都是从物理学领域借用研究方法和工具，如X射线衍射、核磁共振（NMR）、同步辐射。因此，科学工作者不仅需要与同行交流，也要时时关注相关科学领域的进展情况，有哪些理论、方法可以“借用”？有哪些启示可以收集？青年科学工作者心理要有所准备，在重大创新的实践活动中，不乏“十年磨一剑”的例子。在磨剑过程中潜心提高自己科学问题凝练能力，设计解决问题的有效方法，积累重要的科学数据作为攀登的“台阶”和“扶手”。此外，还要对自己的发现不能马上被“大师”们承认的风险有所准备。原始创新有时是孤独的，因此是勇敢者的游戏。

蛋白质结构研究中充满着大量的数学问题，请不要被其吓回头！阅读本书重在设想自己处在科学家们“无助”的位置时，会如何思考遇到的问题，能够有解决问题的不同逻辑吗？请读者与我一起来参加到科学家们的游戏中，体验一下他们的甘苦，重温发现的快乐与激动，也体验一下找不到解决问题出路时“无助”的挫折感，以及重大发现未被承认时的孤独感。你，只有及早体验这些感受，才能及早进入创新的角色中。

本书在介绍科学与技术内容的同时，也尽所能提供一些对蛋白质结构研究科学思想发展的节点和重要手段作出重要贡献的科学家的简短介绍。近不惑之年才刚刚悟出这样的道理：蛋白质结构研究手段及其科学思想发展都是本领域的科学家接力完成的。只有具备了某些必要的素质、品质和思维能力，才能接过前人思想的接力棒，进一步设计新的探索方向，提出科学假说，设计新的实验方法，建立新层次的科学思想和创新理论，才能将科学事业推到更高的境界。为什么只有少数人能够成为杰出科学家，为什么这些人才能够传承顶级的科学思想？正如“钱学森之问”：为什么我们的学校总是培养不出杰出人才？目前我们也刚有些朦胧的感觉，核心应该是“人”。

最后，建议读者在了解科技内容的同时，更要提醒自己："需要培养什么样的思想、素质和能力才能创造新的科学知识?"创新的主体是"人"，读者每读过一章，可以挑一个自己有亲近感的科学家和具体问题，用互联网提供的便利工具，找出自己的"科学家"解决科学问题的每一个台阶和扶手的原始科学文献，看能否拼出其走过路线上的路标，其判断前进方向的逻辑和其中的精妙。也请读者以本书科学家小传内容为跳板，查阅更多关于杰出先辈们思想和行为的文献，找出自己需要培养什么思想品格，需要积累什么经验，进行什么练习？认识到了这些，并坚持身体力行，选定目标，成为大师是时间迟早的问题。

本书成稿仓促，错误难免，欢迎批评指正。

李荣秀

2011 年 3 月

# 目　录


**第 1 章　导论** …… 1

1.1　生命是什么 …… 1
1.2　生物体的有序结构 …… 1
1.3　生物体功能与结构层次 …… 2
1.4　蛋白质是生命活动的载体 …… 2
1.4.1　蛋白质的功能 …… 2
1.4.2　蛋白质结构 …… 4
1.5　蛋白质结构模拟 …… 5
1.5.1　序列比对 …… 6
1.5.2　蛋白质结构比对和预测 …… 6
参考文献 …… 7

**第 2 章　蛋白质结构研究的重大历史事件** …… 8

2.1　假设＋假设＝诺贝尔奖 …… 8
2.1.1　X 射线：波还是粒子？ …… 8
2.1.2　晶体结构之谜 …… 9
2.1.3　诺贝尔奖培训班 …… 9
2.1.4　一石二鸟的精巧实验设计 …… 10
2.2　父子同心：X 射线晶体学 …… 11
2.3　蛋白质晶体学的兴起 …… 12
2.3.1　蛋白质结构 …… 12
2.3.2　多萝西·霍奇金：蛋白质结构测定时代的起始 …… 12
2.3.3　马克斯·佩鲁茨：蛋白质结构测定技术的成熟 …… 13
2.3.4　相位问题 …… 16
2.3.5　又是相位问题：核糖体结构的解析 …… 18
2.4　核磁共振方法 …… 20
2.4.1　傅里叶变换和多维核磁共振 …… 20
2.4.2　测定蛋白质结构的核磁共振方法 …… 22
2.5　蛋白质结构测定的其他辅助工具 …… 24
2.5.1　实验方法 …… 24

2.5.2 计算及结构显示工具 …… 24
2.6 生物分子、复合物及病毒结构测定 …… 25
2.6.1 重要生物学功能蛋白的测定 …… 25
2.6.2 病毒结构的测定 …… 27
2.7 同步辐射光源及发展趋势 …… 28
参考文献 …… 30

**第 3 章 蛋白质结构基础理论** …… 32

3.1 蛋白质结构形成的本质 …… 32
3.2 分子势能 …… 33
3.2.1 量子力学 …… 33
3.2.2 Born-Oppenheimer 近似 …… 34
3.2.3 势能函数 …… 36
3.3 分子内基团识别的要素 …… 37
3.3.1 分子之间相互作用 …… 37
3.3.2 静电相互作用 …… 38
3.3.3 氢键 …… 39
3.3.4 溶剂或疏水相互作用 …… 39
3.3.5 伦敦类型的相互作用 …… 41
3.3.6 原子核在超短距离时的排斥作用 …… 42
3.4 识别的专一性 …… 42
3.5 基团间识别与接触 …… 44
参考文献 …… 45

**第 4 章 从序列到结构** …… 46

4.1 背景介绍 …… 46
4.2 蛋白质二级结构预测 …… 47
4.2.1 Chou-Fasman 方法 …… 47
4.2.2 GOR 方法 …… 47
4.2.3 机器学习 …… 48
4.3 蛋白质三级结构预测 …… 49
4.3.1 蛋白质结构的从头预测 …… 50
4.3.2 蛋白质结构的同源建模 …… 50
4.3.3 同源建模实例 …… 53
4.4 蛋白质四级结构预测 …… 56
4.5 结束语 …… 57
参考文献 …… 57

**第 5 章　蛋白质结构模拟在蛋白研究中的应用实例** …… 59

5.1　引言 …… 59
5.2　蛋白质结构的预测 …… 60
5.3　分子对接 …… 61
5.4　蛋白结构预测与分子对接在生物研究中的应用 …… 64
5.5　分子动力学模拟在蛋白结构变化中的应用 …… 67
5.5.1　研究背景 …… 68
5.5.2　实验方法 …… 68
5.5.3　实验结果 …… 69
5.5.4　结论 …… 71
5.6　分子动力学模拟研究氨基酸突变对蛋白结构的影响 …… 72
5.6.1　研究背景 …… 72
5.6.2　方法介绍 …… 73
5.6.3　实验结果与分析 …… 74
5.6.4　结论 …… 76
5.7　靶标蛋白与抑制剂的相互作用研究 …… 77
5.7.1　研究背景 …… 77
5.7.2　实验方法 …… 78
5.7.3　体系分子动力学轨迹分析 …… 79
5.7.4　结构变化分析 …… 81
5.7.5　结论 …… 85
5.8　本章总结 …… 85
参考文献 …… 86

**第 6 章　锌指核酸酶的构建** …… 88

6.1　锌指蛋白 …… 88
6.1.1　锌指蛋白的分类 …… 89
6.1.2　锌指蛋白的作用及机制 …… 89
6.1.3　锌指蛋白与免疫性疾病 …… 91
6.2　锌指核酸酶 …… 92
6.2.1　核酸酶 …… 92
6.2.2　锌指核酸酶 …… 93
6.3　锌指核酸酶的设计 …… 94
6.3.1　靶位点的确定 …… 94
6.3.2　ZFN 蛋白序列的设计及优化 …… 95
6.3.3　锌指核酸酶设计举例 …… 95

参考文献 …… 106

**第 7 章　诱导多能干细胞的关键转录因子 Oct4 结构模拟与设计** …… 108

7.1　导读 …… 108
7.2　干细胞 …… 108
7.3　诱导多能干细胞 …… 110
7.4　转录因子 Oct4 …… 112
7.5　转录因子 Oct4 结构解析的意义 …… 113
7.6　蛋白质三维结构研究方法 …… 115
7.7　同源模拟法 …… 117
7.8　转录因子 Oct4 的 $POU_S$ 结构域同源模拟 …… 118
参考文献 …… 123

**第 8 章　蛋白质（酶）进化工程** …… 125

8.1　蛋白质的理性设计 …… 128
8.1.1　蛋白质原有性质的改进 …… 128
8.1.2　蛋白质优良性质的组合 …… 130
8.1.3　蛋白质新催化活性的获得 …… 131
8.2　蛋白质的定向进化 …… 132
8.2.1　定向进化的原理及目的 …… 132
8.2.2　定向进化的应用 …… 133
8.3　蛋白质的半合理设计 …… 134
8.3.1　基于结构的有选择的随机突变 …… 134
8.3.2　随机突变后的定点饱和突变 …… 135
8.3.3　随机突变与定点饱和突变同步 …… 135
8.3.4　计算机辅助的半合理设计 …… 135
8.4　应用实例——超嗜热酰基肽水解酶/酯酶的半理性设计 …… 136
8.4.1　引言 …… 136
8.4.2　实验方法 …… 136
8.4.3　实验结果 …… 140
8.4.4　讨论 …… 151
8.4.5　小结 …… 152
参考文献 …… 153

# 第1章 导 论

李荣秀

## 1.1 生命是什么

20世纪是生命科学大发展的时期。奥地利物理学家薛定谔是波动力学之父，是量子力学集大成者之一。正是他从物理学领域闯入生物学领域，在1944年出版了《生命是什么——活细胞的物理学观》一书。在书中薛定谔提出了一系列大胆的猜想：生命现象用物理学和化学理论进行诠释；基因类似非周期性的晶体或固体；突变是基因分子中的“量子跃迁”引起的；染色体是遗传的密码本；生命是“负熵”驱动的，需要从环境抽取“序”维持生命系统的有序组织……这些十分新奇的观点特别引人入胜。

薛定谔这部书出版以后，一批物理学家投身到生物学和遗传学的研究洪流中，新西兰物理学家威尔金斯（1945年转向）和英国物理学家克里克（1947年或1949年转向）就是其中两位。正是《生命是什么》这部书，使克里克放弃了粒子物理的研究计划，钟情于从未打算涉猎的生物学；也使威尔金斯告别了物理学，热衷于探究生命大分子复杂结构的奥妙。此外，美国生物学家沃森在芝加哥读大学时，就被牢牢地吸引住了，以此为契机，他立志献身于揭开生命遗传的奥秘。此外，卢利亚、查尔加夫、本泽等——也都受到这部书的感染。

## 1.2 生物体的有序结构

经过漫长的岁月，地球上的生命从肉眼看不见的单细胞生物进化成今天的藻类、菌类、植物、动物和人类。其共同的特征包括：生物体和周围环境不断进行着物质交换和能量流动的新陈代谢活动，通过调整自身与环境的关系，调节自身内部的各种机能，从而维持生物体在特定空间和特定时段的稳

定状态；许多生物调节系统都比较复杂，往往兼有正向与反向调节作用。从层次角度看，调节可以在分子水平、细胞水平、组织器官水平以及整体水平进行。从调节方式看，存在着诱导、阻遏后反馈抑制。生物体的一个重要特征是自我复制和生长发育。生物体的自我复制表现在分子层面、细胞层面和个体层面：脱氧核糖核酸（deoxyribonucleic acid，简称 DNA）从一条链复制成两条链，一个细胞分裂成两个细胞，以及生物体的繁殖形成更多的相似个体。一只蚊子能够在空中飞舞，在我们皮肤上能够吸血。如果我们一掌拍下去打个正着，那么它就再不能继续吸血，也不能飞走了。因此生物体另一个重要特征就是严整有序的结构。生物体的各种组成部分不是随机混合在一起的，而是非常精妙有序的严整的组装。生物是多层次的有序结构。细胞、组织、器官、系统和个体，每一个层次中的各个结构单元，都有它们各自特定的结构和功能，它们的协调活动构成了复杂的生命系统。

## 1.3 生物体功能与结构层次

动物体的结构可分为系统、器官、组织和细胞几个层次。多个器官组成的能完成一系列连续性生理机能的体系，称为系统。多种组织有序结合起来，组成具有一定形态并完成一定生理功能的结构，称为器官，例如肺、肠等。组织则是形态相似、功能相同的一群细胞和细胞间质的组合。动物体组织分为上皮组织、结缔组织、神经组织和肌肉组织四种。如运动系统由运动控制器官和运动器官两部分组成；运动控制器官由与运动感觉、运动控制有关的纹状体、小脑与丘脑腹前核、腹外侧核构成，还包括大脑运动区（大脑皮质 4 区、6 区）。运动器官由骨骼和骨骼肌组成，骨由骨细胞，胶原纤维及骨基质构成，外被以骨膜或软骨，内有骨髓。每块骨都是一个器官，具有一定形态，具有丰富的血管神经和淋巴管。骨以其分布部位不同分为躯干骨、颅骨和四肢骨。骨骼由骨与骨相互连接成，对人体起支柱、运动和保护作用；各骨之间连接构成骨骼（如颅保护脑，胸廓保护心、肺等。）骨骼肌跨过一个或多个关节，附着在两块或两块以上的骨面上，在神经系统支配下，骨骼肌收缩，牵拉骨骼，产生各种运动。

## 1.4 蛋白质是生命活动的载体

### 1.4.1 蛋白质的功能

地球上所有生命都依赖蛋白质的存在。蛋白质加快人体化学反应的速度，蛋白质使我们能够看、听、尝、闻、感觉、体验、思考和行动。蛋白质

是细胞生命活动的基础。细胞中，除水分外，蛋白质约占细胞内物质的80%，因此构成机体组织、器官的成分是蛋白质最重要的生理功能。动物体各种组织细胞的蛋白质始终在不断更新，如血浆蛋白质的平均半衰期约为10天，肝中大部分蛋白质的半衰期为1～8天，半衰期更短的一些蛋白质只有数秒。成年动物体内每天约有3%的蛋白质更新，借此完成组织的修复更新，维持机体健康与相对稳定。

蛋白质种类繁多，进化越高级的生物体结构越复杂，其蛋白质种类和功能越繁多，以功能不同可以分为以下几类：

（1）结构蛋白　许多蛋白质起支持作用，给生物结构以强度及保护，如韧带含弹性蛋白，具有双向抗拉强度，能够提供机械支撑作用。细胞骨架包含蛋白纤维的复杂网络。在比较大的尺度上，肌肉的收缩依靠大的蛋白聚集体。在其他有机材料中，如头发和骨骼，蛋白质也是重要的成分。所有多细胞生物体的所有组织都含有胶原蛋白，它是脊椎动物最丰富的蛋白，大约占哺乳动物蛋白总量的1/4。

（2）收缩或运动功能蛋白　某些蛋白质赋予细胞与器官收缩的能力，可以使其改变形状或运动。如骨骼肌收缩靠肌动蛋白和肌球蛋白。

（3）营养和储存蛋白　各种离子、小分子和其他代谢物通过与蛋白质结合得到储存，如血红蛋白结合氧、转铁蛋白结合铁。

（4）调节蛋白　合成蛋白质所必需的信息储存于基因（脱氧核糖核酸的序列）中。细胞活动的精密协调需要各种基因产物（蛋白质）在正确的时刻具有适当的量（浓度）。酶通过翻译DNA序列合成蛋白质，这个生产过程又能被其他蛋白以复杂的机制促进和抑制。

（5）运输蛋白　蛋白质参与从电子到大分子的运输。铁由转铁蛋白运输、血红细胞中的血红蛋白从肺运输氧到其他组织，并把二氧化碳运回到肺。蛋白质在膜上形成通道运输离子，即使蛋白质的跨膜运输也需要依赖其他的运输蛋白。

（6）催化蛋白　生物体系中的化学反应主要由酶（一种蛋白质）来催化。从简单的二氧化碳水合到生物大分子的复杂转化，都需要由酶进行催化才能完成。酶参与了从DNA读取遗传信息到合成细胞中的物质的一系列化学反应，是细胞机器的核心。它可使反应速度加快数百万倍。

（7）信使蛋白　蛋白质通过作为神经细胞间小分子的受体参与神经信号的传递。生物体内不同器官、组织和细胞中的生物过程必须得到有效的协调。能起协调作用的是激素。部分激素是蛋白质，同时，激素受体也是蛋白质，能起到传递信号的作用。

（8）防御蛋白　如抗体是由免疫系统产生，能够结合外源颗粒（如病毒和细菌）的蛋白质。

(9) 其他蛋白　如病毒和噬菌体是核蛋白，病毒可以致病。这些生物功能存在于生物体繁殖、生长、结构和死亡过程的各个方面都依靠生物识别和结合的精密机制。在分子水平上，从基因的复制、转录和表达到基因的调控，从生物催化剂——酶的活性到激素的功能，以及有机体对外源有害物质如毒素、病毒、细菌的抵御和清除过程，无不充满了这种机制。细胞对接触到的每一种外源物质进行辨别；细胞内的几乎每一种物质都首先被代谢它的酶或细胞表面的受体结合，从而吸收进入细胞。蛋白质之间可以紧密结合形成复合物，蛋白质和核酸结合控制基因的复制与表达，蛋白质和细胞表面的多糖结合，蛋白质和脂类结合整合到细胞膜上。这些生物功能的物质基础是蛋白质结构。

### 1.4.2　蛋白质结构

蛋白质是具有特定长度的线性非均一聚合物，如某种蛋白质包含的单体的种类和组成是恒定的，而不同蛋白质则包含不同种类和数量的单体，从几个到几千个不等。线性链再折叠成为三维构象。蛋白质构象的性质由单体（氨基酸）的序列决定，因此蛋白质的可能序列具有非常显著的差异，相应地，蛋白质的三级结构也具有极大的多样性，从纤维状到球状。蛋白质的结构一般分为几个层次：

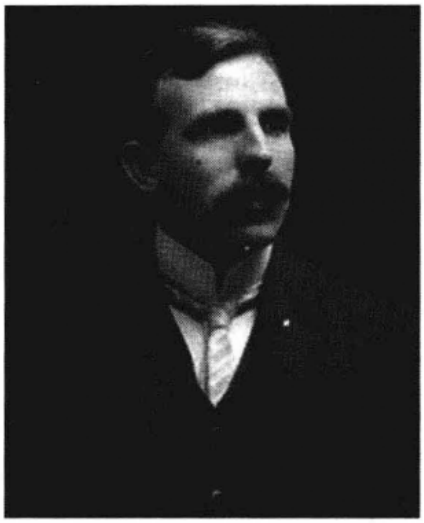


图 1.1　恩内斯特·卢瑟福

(1) 一级结构　构成蛋白质的单元氨基酸通过肽键连接形成的线性序列，为多肽链。

(2) 二级结构　一级结构中部分肽链的弯曲或折叠产生二级结构。多肽链的某些部分氨基酸残基周期性空间排列。卷曲所形成的二级结构称为 α-螺旋，折叠所形成的二级结构称为折叠片。这两种二级结构的形成都是由于距离一定的—N—H 基团和—C═O 基团之间形成氢键。

(3) 三级结构　在二级结构基础上进一步折叠成紧密的三维形式。三维结构一般都是球状或纤维状的。

(4) 四级结构　由蛋白质亚基结构形成的多于一条多肽链的蛋白质分子的空间排列。

恩内斯特·卢瑟福（Lord Ernest Rutherford）（图 1.1），现代原子科学的奠基人，原子之父。生于新西兰 Nelson 附近的泉林村，父亲是农民，母亲是乡村教师。1895 年考取大英博览会奖学金，到英国剑桥大学卡文迪什实验室学习。1911 年由 α 粒子散射实验推出卢瑟福原子模型，指出原子有一个极小的核，几乎集中了原子的全部质量，而电子则绕核旋转，而非之前

人们认为的“葡萄干布丁”模型（电子“镶嵌”在原子上，就像布丁上的葡萄干）。

卢瑟福是20世纪初最伟大的实验物理学家，1908年诺贝尔化学奖得主，一生发表论文约215篇，著作6部，培养了10位诺贝尔奖获得者。卢瑟福的实验室被后人称为“诺贝尔奖得主的幼儿园”。他的头像出现在新西兰货币的最大面值——100元上面，作为国家对他最崇高的敬意和纪念。

**小提示：别忘了思考！**

英国剑桥大学的卡文迪什实验室，一直坚持这样的规定：每天下午6点整，老资格的研究人员来到实验室，宣布时间已到，要求每个人停止工作。如果谁不遵守，他们便引用Rutherford的话加以劝导。Rutherford说过：“谁未能完成六点前必须完成的工作，也就没有必要拖延下去，倒是希望各位马上回家，好好想想今天做的工作，好好思考明天要做的工作。”那是一天深夜，Rutherford披着外衣，又来到实验室检查，惊奇地发现有人还在做实验。由于低头，又十分专心，那学生没发现Rutherford站在他的身后。Rutherford轻声地问道：“你上午干什么?”学生回头一看，是Rutherford，他马上站起来，小心地回答：“做实验。”Rutherford又问：“那么，下午呢?”学生回答：“做实验。”Rutherford提高了声调，再问：“晚上呢?”学生以为老师在表扬他，得意地回答：“还是做实验。”Rutherford极为严肃地问：“你整天做实验，还有时间去认真思考吗?!”那学生低下了头。临走，Rutherford告诫他：“别忘了思考!”从此，卡文迪什实验室的人记住了Rutherford的忠告：“别忘了思考!”

## 1.5 蛋白质结构模拟

随着人类基因组、植物基因组和微生物基因组等大规模研究计划的进展，基因数据库每年可增加数亿个碱基序列，数据库的大小每14个月翻一番。UniProtKB/TrEMBL（unreviewed，automatically annotated）蛋白质数据库2009年7月7日公布的40.5版包括8594382个条目。这些蛋白质序列主要从公开的基因和核酸序列数据库（如EMBL-Bank/GenBank/DDBJ database）直接翻译编码序列而成。“UniProtKB/Swiss-Prot”数据库用手工标注，加上从文献收集的信息，再进一步汇集成库，2009年7月7日公布的57.5版包括470369个条目。虽然X射线衍射技术和NMR技术近年有了很大发展，使越来越多的生物大分子高级结构得到了解决，但Brookhaven蛋白质结构数据库（Protein Databank，PDB）到2009年7月

21 日也只有 59027 个晶体结构。

随着结构生物学的发展，越来越多的生物大分子晶体结构得以发表，提供了许多生物大分子功能的结构基础。生物体的信号传递过程和功能性相互作用的机理也越来越清楚，计算机中央处理器（CPU）速度更快，功能更强的分子模拟及分子设计软件的发展，使得在分子水平设计并合成物质模拟生物大分子成为可能。

### 1.5.1 序列比对

序列比对（sequence alignment）的基本问题是比较两个或两个以上 DNA 和蛋白质序列的相似性或差异性。从生物学的初衷来看，这一问题包含了以下几个意义：①从相互重叠的序列片段中重构 DNA 的完整序列，将各种试验条件下的探测数据（probe data）构建为基因完整的物理图谱；②通过序列比对发现生物序列中的功能、结构和进化的信息；③生物学关系密切的基因和蛋白质都具有一些不连续的共同序列，序列比对发现的局部序列高度相似比完整序列的全局比对更有意义。序列比对的方法包括全局对齐、局部对齐、空隙惩罚等，序列比对结果中有匹配、不匹配、插入和删除以及空隙几种区域；两个序列比对常采用动态规划算法，适用于长度较小的序列。而对于海量基因序列（如人的 DNA 序列高达 $10^9$bp）就不太适用，甚至采用复杂的非线性算法也难以奏效。因此，启发式方法的引入势在必行，著名的 BALST 和 FASTA 算法及相应的改进方法均是在这种思想上发展起来的。

### 1.5.2 蛋白质结构比对和预测

基本问题是比较两个或两个以上蛋白质分子空间结构的相似性或不相似性。蛋白质的结构与功能是密切相关的，一般认为，具有相似功能的蛋白质结构是相似的。蛋白质是由氨基酸组成的长链，长度从 50AA 到 1000～3000AA（amino acids）。蛋白质具有多种功能，如酶催化功能、物质的存储和运输、信号传递、抗体功能等。蛋白质的氨基酸序列决定了其三级结构。一般认为，蛋白质有四级结构。研究蛋白质结构和预测的理由是：可以理解生物的功能；寻找 dockingdrugs 的目标；获得更好的农作物的基因；合成酶应用于工业；直接对蛋白质结构进行比对的原因是由于蛋白质的三级结构比其一级结构在进化中更保守，同时也包含了较 AA 序列更多的信息。蛋白质三级结构研究的前提是假设内在的氨基酸序列与三级结构一一对应（不一定全真），物理上可用最小能量来解释。从观察和总结已知结构的蛋白质结构规律出发来预测未知蛋白质的结构，同源建模（homology modeling）和指认（threading）方法属于这一范畴。同源建模用于寻找具有高度相似性

的蛋白质结构（超过30%氨基酸相同）；后者则用于比较进化族中不同的蛋白质结构。然而，蛋白质结构预测研究现状还远远不能满足实际需要。

（1）药物　在一定程度上，可以设计蛋白质，如利用蛋白质工程改造胰岛素，使胰岛素更容易在水中溶解成为活性形式，从而更快地发挥效果。了解酶的功能，使人们能够设计药物，抑制特定的目标酶的活性。

（2）农业　蛋白质结构和功能的知识也能够用于治疗植物病害，调节植物的生长和发育，如培育储存时间更长的水果。

（3）工业　蛋白质工程具有设计合成催化各种大规模工业过程用酶的潜力。如有科学家正在研究利用脂酶工业化水解脂肪。

## 参考文献


[1] Noji H, et al. Nature, 1997, 386: 299-302.

[2] Stupp S I & Braun P V, Science, 1997, 277: 1242-1248.

[3] 蛋白质三级结构数据库 http://www.rcsb.org/pdb/

[4] 蛋白质序列数据库 http://www.expasy.ch/sprot/sprot-top.html

# 第2章 蛋白质结构研究的重大历史事件

李荣秀

## 2.1 假设+假设=诺贝尔奖

### 2.1.1 X射线：波还是粒子？

德国慕尼黑大学的物理学家威尔姆·康拉德·伦琴（Wilhelm Conrad Röntgen）1895年发现了X射线（又称伦琴射线，Röntgen rays；X射线），因此而获得了1901年度的诺贝尔物理学奖。X射线的发现开创了人类探索物质世界的新纪元，揭开了20世纪物理学革命的序幕。X射线的发现不仅激起公众对其医学影像应用的遐想，也激起物理学家们对其本身是什么的争论。当时物理学正处于革命性过渡的中期，光是一种波还是一种粒子的争论已经持续了250多年。伦琴发现X射线5年后，马克斯·普朗克（马克斯·卡尔·恩斯特·路德维希·普朗克，Max Karl Ernst Ludwig Planck，德国物理学家，量子力学的创始人，20世纪最重要的物理学家之一，因发现能量量子而对物理学的进展做出了重要贡献，获得1918年诺贝尔物理学奖）提出了量子理论，认为像物质由原子组成一样，能量是由称为“量子”（quanta）的粒子组成。1905年爱因斯坦（阿尔伯特·爱因斯坦，Albert Einstein，著名的犹太裔理论物理学家、思想家及哲学家，也是相对论的创立者，是现代物理之父及20世纪最重要的科学家之一）发展了普朗克的理论，证明光电效应（photoelectric effect），光照射到金属表面使电子脱出的现象，只能用光的粒子理论解释。尽管10年后人们最终认识到光既有波动性又有粒子性，但当时并没有定论，有待于科学证明。当时，伦琴已经部分证明了X射线的波动性质，如在照相底片上有效应，但在从一种物质进入另一种物质时没有折射现象，也不能被反射。X射线不能被强磁场折射，因此不太可能是带电粒子。但爱因斯坦的观点仍存在极大的影响，“如果X射

线是光，那么金属块怎么可能像一个家庭主妇那样收集储存 X 射线波的碎片，直到自己的一个儿辈电子身上的能量足够大，可以穿越真空产生 X 射线?”1904 年，英国利物浦大学的查尔斯·格洛弗·巴克拉（Charles Barkla，英国物理学家。任教于剑桥大学、爱丁堡大学。因发现 X 射线的散射现象，获得 1917 年诺贝尔物理学奖）证明 X 射线能够被极化。尽管爱因斯坦强烈反对，但极化证据提示 X 射线是横波。进一步的证据表明，X 射线是电磁波，波长只有可见光波长的万分之一。阿诺德·约翰内斯·威尔海姆·索末菲（Arnold Johannes Wilhelm Sommerfeld，德国物理学家，量子力学与原子物理学的开山始祖之一）在 1912 年《物理学年鉴》（德文 Annalen der Physik，英文 Annals of Physics）的一篇论文中写道，“衍射证据可能是 X 射线波动理论的拱心石”，关键要看 X 射线是否存在衍射或干涉效应。

### 2.1.2 晶体结构之谜

冯·劳厄（马克斯·冯·劳厄，Max von Laue，德国物理学家，因发现晶体中 X 射线的衍射现象而获得 1914 年诺贝尔物理学奖）从中学时期就喜爱光学，尤其是光的波动理论，因此他不假思索就接受了索末菲撰写《数学科学百科全书》中“波动光学”条目的提议，这让他有机会思考晶体结构的晶格理论。奥古斯特·布拉菲（Auguste Bravais，法国物理学家）1850 年曾提出晶体是由原子非常规则的排列形成在三维空间重复的几何形状，即点阵结构。在晶体学家冯·格罗特（Paul von Groth）的支持下，慕尼黑大学的许多系所都陈列有晶体点阵结构模型，但当时并没有得到实验证实。

### 2.1.3 诺贝尔奖培训班

1914～1918 年间，慕尼黑是科学发现和艺术进步的世界中心。绘画、诗歌和物理学在德国巴伐利亚州首府极其繁荣。主要人物们经常光顾的聚会场所是位于霍夫花园（Hofgarten）的咖啡店。正如艾略特（T. S. Eliot，1948 年诺贝尔文学奖获得者）在“不朽的荒原”（The Waste Land）中描写的那样：

Summer surprised us，coming over the Starnbergersee.

（夏天令我们吃惊，它越过斯塔恩贝格湖。）

With a shower of rain；we stopped in the colonnade.

（带来阵雨；我们躲进柱廊里。）

And went on in sunlight，into the Hofgarten.

（天放晴后再出来，进入霍夫花园。）

And drank coffee，and talked for an hour.

（在那里喝咖啡，谈了一小时。）

霍夫花园咖啡店的咖啡和谈话不仅仅是诗人们的特权，也为来自附近不同大学的物理学家们提供了一个非正式讨论交流的好场所。这些讨论在发现X射线本质的问题上将起到重要作用。这些咖啡会议的一个核心人物就是索末菲。索末菲是具有良好数学基础训练的理论物理学教授，为了弥补物理学实验知识的不足，他请伦琴的助手亚伯拉罕·乔夫（Abraham Joffe）对其进行一些训练。乔夫建议每天早晨与索末菲在霍夫花园咖啡店会面交流实验物理学。索末菲开始是勤奋的学生，也是才华出众的老师，这个非正规会面很快扩大到晶体学家和化学家，与周三晚上大学的研讨会一起创建出一种当时很特别的开放、创造性学术氛围。索末菲一生得过无数的荣誉，如马克斯·普朗克奖章、洛仑兹奖章（Lorentz Medal）、奥斯特奖章（Oersted Medal），选入皇家学会、美国国家科学院、俄罗斯科学院、印度科学院，以及许多其他在柏林、慕尼黑、格丁根、威尼斯等地的科学院，拥有许多大学颁发的荣誉博士学位，唯一没有得到的是诺贝尔奖。索末菲知道如何发掘与招募优秀人才，提升他们对物理与数学的兴趣，培育他们成为理论物理的精英。索末菲的这个“索末菲学校”有多达7位校友获得了诺贝尔奖，包括沃尔纳·卡尔·海森堡（Werner Heisenberg，1932年获诺贝尔物理学奖）、彼得·约瑟夫·威廉·德拜（Peter Joseph William Debye，1936年获诺贝尔化学奖）、伊西多·艾萨克·拉比（Isidor Isaac Rabi，美国物理学家，1944年获诺贝尔物理学奖）、沃尔夫冈·泡利（Wolfgang Pauli，奥地利物理学家，1945年获诺贝尔物理学奖）、汉斯·贝特（Hans Bethe，美国物理学家，1967年获诺贝尔物理学奖），以及马克思·冯·劳厄。冯·劳厄于1909年作为私人教师加入“索末菲学校”，马上就被这里“充满X射线本质的问题”深深触动。

### 2.1.4 一石二鸟的精巧实验设计

彼得·保罗·埃瓦德在进行关于长电磁波在假定的晶体空间点阵上的行为研究中遇到一些困难，于1912年2月的一个晚上来征求劳厄的建议。劳厄无法回答埃瓦德的问题，但他的思绪开始遐想。他头脑中突然出现了一个联系：既然光线会在缝隙与光波长相近的光栅上产生衍射和干涉，那么如果X射线是一种波的话，其波长又是可见光波长的万分之一，那么晶体原子（离子）的间距可能正好可以用作X射线的天然光栅。如果这个推论成立的话，那么当用X射线照射晶体时就应能在照相底片上观察到典型的干涉亮点图案。在一次滑雪中，劳厄和索末菲、维恩（W. Wien）及其他同事谈到这个想法时，得到了全票质疑，怀疑实验的可行性，但索末菲仍然大方地允许劳厄进行这项探索。劳厄设计了实验，在X射线管和照相底片间放置一块硫酸铜晶体。索末菲的助手弗里德里希（Walter Friedrich）和尼平（Paul

Knipping）开始了实验探索，在经历了起初的一些无结果的实验后，于1912年4月23日终于成功！X射线穿越晶体在照相底片上形成了亮点图案，证明了推断的正确性。在一个精美的实验中，劳厄同时证明了X射线的波动性质和晶体结构的空间点阵性质。当劳厄接受诺贝尔奖时，他衷心感谢弗里德里希和尼平在这项发现中的贡献，并与他们分享了奖金。这是固体物理学研究中具有里程碑意义的发现，从此，人们可以通过观察衍射图案研究晶体的微观结构，并且对生物学、化学、材料科学的发展都起到了巨大的推动作用。

晶体的点阵结构在当时虽然还是一种假设，X射线波动理论正好有待于证明，这两件事通过讨论晶体光学问题“碰巧”在劳厄的脑子里形成了联系！他敏感地抓住了晶格间距的数量级和X射线波长可能相近的可能性，从而可起到X射线天然光栅的作用。他的灵感来自他同时对这两个互不相干的假设都有明确具体的概念。如果不知道晶体的点阵结构的假设，也不知道X射线的波动理论，他永远也不会想到推断出利用晶体进行X射线衍射的实验设计。

## 2.2 父子同心：X射线晶体学

马克思·冯·劳厄发现X射线晶体衍射的现象为测定晶体结构奠定了坚实的基础。两个英国科学家，威廉·亨利·布喇格爵士（Sir William Henry Bragg）和威廉·劳伦斯·布喇格爵士（Sir William Lawrence Bragg）开始用X射线衍射测定化合物中原子在晶体中的位置。

亨利·布喇格发明了X射线分光计，并与他的儿子威廉·劳伦斯·布喇格创立了用X射线分析晶体结构的新学术领域。

劳伦斯·布喇格1912年就发现了其一生最著名的成就——布喇格定律。布喇格定律给出了受到电磁辐射和粒子波照射时，晶体内原子平面间隔与在该平面上产生最强反射的入射角之间的关系，可以用于测定波长和晶体的点阵间隔距离。其间，他与在利兹大学研究X射线分光计的父亲进行了讨论。父子两人对X射线衍射的研究为后来DNA双螺旋结构的发现奠定了基础。

1937年欧内斯特·卢瑟福去世后，劳伦斯·布喇格接替他成为剑桥大学卡文迪什实验室主任。1948年，劳伦斯·布喇格开始对蛋白质结构研究产生兴趣，并且创建了一个使用物理方法解决生物问题的研究小组。1953年DNA双螺旋结构的发现，他尽管没有直接参与研究，但也做出了贡献，在他的支持下，弗朗西斯·克里克（Francis Harry Compton Crick）和詹姆斯·沃森（James Dewey Watson）在卡文迪什实验室开展这方面的研究。而他40年前做的X射线研究也在这一过程中得到应用。同一时期，马克

斯·佩鲁茨在卡文迪什实验室开展血红蛋白研究，后来他也因这项研究获得1962年诺贝尔化学奖。经劳伦斯·布喇格提名，克里克、沃森以及莫里斯·威尔金斯成为1962年诺贝尔生理学或医学奖的候选，并最终获奖。

布喇格父子创立了用X射线分析晶体结构的新领域，这项技术的应用为DNA双螺旋结构的发现和蛋白质晶体结构的测定奠定了基础。正是由于这项成就，1915年父子两人一同被授予诺贝尔物理学奖。

## 2.3 蛋白质晶体学的兴起

### 2.3.1 蛋白质结构

生物大分子相互作用是各种生命现象和生物功能的物质基础和本质。如果把核酸（DNA、RNA）比喻为指导细胞内各种活动的剧本，那么蛋白质就既是导演又是“主要演员”。蛋白质不断与核酸相互作用（读取剧本），按照剧本创造作为主要演员的蛋白质；蛋白质还不断与蛋白质相互作用，调节蛋白质生物活性、功能和蛋白质的归宿。这些相互作用都依赖于这些生物大分子本身的空间结构。

自然界中蛋白质主要是由20种不同的氨基酸通过化学键连接成一个长链大分子，链上氨基酸的排列顺序为蛋白质分子的一级结构。每种蛋白质分子都有其独特的一级结构，具有生物活性的蛋白质并不以笔直或随意弯曲的飘带形式存在，而是出现其固定的空间结构（三级结构），有一定规律性的局部结构，如α-螺旋、β-折叠等，称为蛋白质的二级结构。有两个或更多相对独立的链可以形成更复杂的空间结构。蛋白质分子的一级结构决定了其空间结构，正常情况下，氨基酸序列相同的蛋白质分子具有相同的空间结构。而蛋白质分子相互作用的活性除与氨基酸序列有关之外，还依赖于整个分子的空间结构。有正常的氨基酸序列、没有形成特定的空间结构的蛋白质分子是没有生物活性的，这称为失活的蛋白质分子。要了解生物结构现象，了解蛋白质晶体结构只是非常重要的第一步。因此，研究生物大分子的空间结构一直是整个生命科学中最重要、最活跃的领域之一。

### 2.3.2 多萝西·霍奇金：蛋白质结构测定时代的起始

多萝西·霍奇金（Dorothy Hodgkin，1964年诺贝尔化学奖得主）1928年进入英国牛津大学萨默维尔学院（Somerville College）攻读化学期间，被晶体之美深深吸引，于1932年到英国剑桥大学师从贝尔纳（J. D. Bernal，英国科学家，一生在结晶学、分子生物学、科学社会学等方面有过许多天才的思想）攻读博士学位。正是在贝尔纳的实验室，霍奇金第一次用X射线

技术研究生物分子，用她自己的话讲这是“一个充满新发现的时期”。1934年，他们收到瑞典乌普萨拉大学斯维德伯格实验室（laboratory of Theodor Svedberg，Uppsala University，Sweden）的约翰·菲尔波特（John Philpott）制备的胃蛋白酶晶体。贝尔纳发现维持蛋白晶体不干燥的重要性，而且只有用其母液保存才能得到清晰的X射线衍射点。霍奇金努力鉴定这个新晶体。从她的X射线衍射照片，他们推测这种酶由直径3.5nm和2.5nm扁椭球体分子的六方晶格组成，同时晶格中还含大量共结晶水，最重要的是，从衍射点强度他们能够推断其中的原子也具有完美的排列。“从结晶的蛋白质可以制取X射线底片，我们已经具备了查看所有结晶蛋白结构的工具，得到的蛋白质结构细节比物理和化学方法精细得多”。1934年的这篇论文标志着蛋白晶体学的开端，进入了一个新的纪元。在近年的高度竞争的压力下，可能再也难见到这样科学研究合作的成功案例了。

格兰·弥黎荻博士是一个年轻的美国人，其父亲是诺贝尔奖获得者罗伯特·弥黎荻（Robert Millikan），很熟悉贝尔纳的研究。他看到胃蛋白酶晶体的第一个直觉就是贝尔纳会很高兴得到这些晶体，于是就拿到了剑桥。霍奇金和贝尔纳用实验证明X射线衍射强度的明显变化不是失去生物活性的结果，而是晶格变化的结果。霍奇金进行蛋白质结晶实验开创了蛋白质结构测定的时代，因而获得了1964年度的诺贝尔化学奖。正如佩鲁茨所写的那样，霍奇金建立的X射线晶体学成为一种进行天然物质化学组成测定最快的方法。人们普遍认为，霍奇金在20世纪30年代成功拍摄的胃蛋白酶和胰岛素X射线衍射照片是蛋白质晶体学研究历史的里程碑。

### 2.3.3 马克斯·佩鲁茨：蛋白质结构测定技术的成熟

马克斯·佩鲁茨（Max Ferdinand Perutz）1932年进入维也纳大学，用他自己的话说是“在无机分析课程上浪费了5个学期”。凡·卫斯理（F. von Wessely）教授的课程启发了他对有机化学（尤其是有机生物化学）的好奇心。1936年9月佩鲁茨成为剑桥大学卡文迪什实验室的一名研究学生，指导老师是贝尔纳。此后，佩鲁茨就没有离开过剑桥。希特勒入侵奥地利和捷克斯洛伐克后，佩鲁茨家族的生意被没收，他的双亲流落为难民，他自己的经济支持很快耗尽。好在劳伦斯·布喇格爵士（Sir Lawrence Bragg）用从洛克菲勒基金会得到的一笔研究经费资助他，1939年1月1日任命他为研究助理。由于战争的影响，经费时断时续地维持到1945年，佩鲁茨才争取到一笔帝国化学工业研究奖学金。1947年他受聘于新建立的医学研究委员会（Medical Research Council）研究分子生物学，全体员工除了他只有约翰·考德雷·肯德鲁（John Cowdery Kendrew）一个人。他一直待到1962年3月被任命为医学研究委员会分子生物学实验室（Medical Research

Council Laboratory of Molecular Biology）主任。这些年他一直坚持与劳伦斯·布喇格爵士进行合作研究。

1938年初，贝尔纳、范库谢和佩鲁茨发表了血红蛋白和chymotrypsin晶体X射线衍射的合作研究论文［Nature，1938（141）：523］。糜蛋白酶晶体是双晶，难以研究，佩鲁茨就选择血红蛋白继续研究。生物与寄生虫学教授凯林（D. Keilin，剑桥大学）很快对这项工作产生了兴趣，为佩鲁茨和同伴提供了生物化学研究设施，而这些正是卡文迪什实验室所缺乏的。因此，从1938年开始到50年代早期，蛋白质化学的工作一直在凯林的莫尔蒂诺研究所（Molteno Institute）进行，X射线方面的工作在卡文迪什实验室进行。佩鲁茨对那些天才同事能够加入一起工作心存感激，其中肯德鲁1946年加入，弗朗西斯·克里克1948年加入，詹姆斯·沃森1951年加入，弗雷德里克·桑格（Frederick Sanger）1958年加盟。他的这些同事都先后获得了诺贝尔奖。肯德鲁和佩鲁茨（Max Ferdinand Perutz）1937年开始氧储存蛋白-肌红蛋白（myoglobin）高分辨率结构的研究工作，于1958年发表论文，随后又测定了氧输运蛋白-血红蛋白（hemoglobin）的高分辨率结构。肯德鲁和佩鲁茨赢得了1962年诺贝尔化学奖。蛋白质结构与克里克和沃森早5年发表的脱氧核糖核酸（deoxyribonucleic acid，DNA）结构相比，极不规则，表现出形状多样的特点。这种结构多样性与蛋白质功能多样性的关系在随后的时间里得到了更多的证明。

1958年开始，佩鲁茨指导罗斯曼（Michael Rossmann）研究血红蛋白结构时发展了分子置换技术（molecular replacement），可以已知结构推断相关但未知结构的取向和位置。

化学家的重要工作之一是用他们自己的方式忠实复制化合物的结构，其中原子是建筑材料，总原子个数、相互之间的连接点和连接方式即为原子在空间的分布。结构测定一直是化学研究的最重要工作，测定的方法也很多。由于多种原因，有机化合物结构测定曾经发展很迅速，基本用理论化学就能够满足需要，可以从化合物的反应性质、降解产物以及用简单物质进行合成验证，就可以推测出原子间相互连接的一般化学结构，而不是原子间精确距离和键间夹角。然而，要建立结构与性质之间的关系，原子间精确距离和键间夹角是必不可少的，只有通过物理方法才能测定。

X射线遇到晶体时产生向特定角度的衍射，衍射的角度、强度与晶体中原子的种类和空间发布相关，因此可以进行结构测定。马克思·冯·劳厄因发现这个衍射现象获1914诺贝尔物理奖，开辟了研究X射线本质和固相化合物结构的全新领域。布喇格父子1915年首先发展出了测定结构的方法，并因此也获得了诺贝尔物理奖。

测定方法不断改进，能够测定越来越复杂的结构，但当遇到不那么简单

的结构时，就遇到了相当大的困难。从实验数据到化合物结构之间没有通用的简单方法可用，数学计算特别耗时。尽管这样，到 20 世纪 40 年代中期，X 射线衍射能够测定的分子结构比其他经典化学方法测定的都复杂得多。

1937 年还没有测定蛋白质结构的方法，佩鲁茨在剑桥实验探索用 X 射线衍射测定血红蛋白结构的可能性。劳伦斯·布喇格爵士不知疲倦地继续开展和他父亲共同开创的事业，1938 年成为卡文迪什实验室的主任。当他看到佩鲁茨获得的实验数据时，对佩鲁茨加以鼓励，并提供了很有效的支持。血红蛋白是在生命过程中起重要作用的一类分子，是生物的基本成分，是红细胞的功能成分。血红蛋白含的铁在肺中结合氧，然后在身体其他组织中释放出来。血红蛋白是球蛋白的一类，分子形状近乎球形。一开始就选择它进行研究，一方面是因为正好能够得到这个分子的完美结晶，另一方面对于蛋白质来说，血红蛋白分子恰好相当小，困难会相对小一些！大约 10 年后肯德鲁加入佩鲁茨研究组后，解析肌红蛋白结构的任务就恰好安排给了肯德鲁。肌红蛋白是另一种球蛋白，与血红蛋白密切相关，而其大小只有血红蛋白分子的 1/4。可以在肌肉中找到肌红蛋白，在那里它起到储存氧的作用。鲸和海豹肌肉组织中的肌红蛋白特别丰富，因为它们需要能够存储大量氧，满足活组织的需要。

然而佩鲁茨和肯德鲁遇到了相当大的困难。尽管他们进行了非常全面的研究工作，直到佩鲁茨成功将重原子（当时是汞原子）吸收到血红蛋白晶体中特定位置，理想的研究结果才姗姗来迟。这样衍射图案会发生某种变化，变化本身能够直接用于结构测定。该方法在当时原则上已经公开，但佩鲁茨在新的高度使用得更加熟练。肯德鲁用另一种方法使肌红蛋白晶体吸收重原子，其处理方法就类似。这种技术的必要前提条件是重原子的掺入不会影响晶体中其他原子的位置。就此而论，就是因为分子几何尺寸相对很大，蛋白质分子基本不变。布喇格相当贴切地指出，蛋白质分子对这些附件一点也不在乎，就如印度象不在乎额头上画个金星！

尽管直接测定血红蛋白和肌红蛋白的路线已经走通，仍然有大量数据需要处理。两个分子中较小的肌红蛋白约含 2600 个原子，其中许多原子的空间位置已经知道。这些结果是肯德鲁研究 110 个晶体、测定约 250000 幅 X 射线衍射图的强度才得到的。如果不使用计算速度较高的电子计算机的话，这些计算几乎不可能。血红蛋白分子有肌红蛋白的 4 倍大，对其结构了解更少。这两种情况需要佩鲁茨和肯德鲁收集更多的映像去获得结构的更多细节。

由于佩鲁茨和肯德鲁的贡献，现在知道了球蛋白的结构原理，这是经过 25 年艰难探索才得到的，尽管初始结果并不完美。因此，让我们佩服的不仅仅是两位科学家在这项研究工作中表现出的天才和技巧，更值得我们学习的是他们的耐心和毅力，正是这种人类独有的优秀品质让他们征服了那些开

始时看来不可能逾越的艰难险阻。目前 Brookhaven 蛋白质结构数据库（Protein Databank，PDB）近 6 万个生物分子三维结构数据（2009 年 7 月 21 日）的大部分都是用佩鲁茨和肯德鲁奠定的 X 射线衍射方法测定解析的。今天我们能够更加确切地理解这些生命必需物质结构知识的发现和积累对于阐明生命现象本质的重要性，也更能够清晰认识诺贝尔设立的获奖条件具体实例的阐述。佩鲁茨和肯德鲁的贡献为进一步在分子水平理解生物功能的结构基础打开了一条通往新世界的大门。

### 2.3.4 相位问题

对于佩鲁茨来说，证明 X 射线能够解析生物学上重要的蛋白质复杂结构不仅需要灵感，更需要毅力。

当 X 射线穿过晶体时，常常被原子折射（deflect），再相互影响，生成典型的 scattered 图案，光点之间距离也极有规律，这就是 X 射线衍射图案（X-ray diffraction pattern）。布喇格和他的父亲首先发现这些衍射图案能够揭示化合物结构，但当科学家用布喇格的 X 射线晶体学方法研究越来越复杂的结构时，很快就发现了方法适用的局限性！这些 X 射线波碰到膜后形成衍射图案中的斑点。每个衍射点提供的信息只有晶体中原子偏射该点方向 X 射线波的振幅，而理想情况下科学家也想知道在波动周期运动中一个峰到下一个峰上的每个点，即波的相位。对于只有 10 多个原子的分子来说，用 19 世纪法国数学家傅里叶（Joseph Fourier）设计的将复合波分解成单波的公式（即傅里叶变换方法）就可以从衍射图案和波的振幅数据推算出分子结构。尽管推算过程极其艰辛枯燥、耗时费力，早期的晶体学科学家几乎总是能够依靠这个傅里叶公式，结合直觉和一定次数的尝试，找到正确的分子结构。

而当遇到像血红蛋白分子这样含数千个原子的大分子时，出现了更大的困难。这么大分子的晶体中反射和相互作用的数量如此大、如此复杂，导致 X 射线衍射图案转化为分子结构需要更多的数据，而仅仅有波的振幅数据是不够的。需要找到一个方法确定衍射图案中每个衍射点形成时 X 射线波的相位（phase），推算大分子结构。如果知道相角、波上点与特定参照点的间距，就可以计算出相位。但实际上用傅里叶变换方法或其他任何已知方法都不能直接计算出相位。

科学家称这种相位信息缺失的情况为“相位问题”（the phase problem），在当时，解决这个问题成为致力于用 X 射线晶体学测定蛋白质结构的科学家们的终极目标。在缺失相位数据的情况下，科学家求助于另一个可以绕过这个问题的方法。

帕特森（Lindo Patterson）在 1934 年就设计出了傅里叶变换的函数，

只用X射线衍射图案中点的位置和强度就能够画出等场强线图，定义晶体中原子间的距离。帕特森图的优点是没有反射的相位数据也能解决问题，缺点是即使对于简单的化合物，画帕特森图也是一个极其耗时的辛苦劳作，衍射图上每个点都需要进行数百次计算。此外，用帕特森图也不能保证对复杂化合物能够处理成功。

在没有其他方法可选用的情况下，佩鲁茨只得创作血红蛋白的帕特森图，他自己后来称其为他自己领域中最枯燥和最差的工作之一。

克里克认为有一个方法可能会成功解析出蛋白质的结构，那就是格拉斯哥大学罗伯逊（J. M. Robertson）研究有机小分子所用的同晶替换方法。在不扰动分子形状的前提下向晶体中引入比分子中原子重的原子，重原子携带更多的电子，散射X射线的方式不同。含重原子和不含重原子的两幅衍射图的点基本相同，强度不同的关键点能够提供给科学家精确定点重原子的位置。从这些重原子位置可计算反射X射线的相位。当时没有人能肯定同晶置换方法在蛋白质结构研究方面是否行得通。血红蛋白含数千个原子，一个重原子的作用有多大?

佩鲁茨推理到引进重原子的方法看来更不可行。撇开化学上的困难不谈，从重原子散射出的强度是否大到能变更像蛋白质这样大的分子的绕射图形，仍是个问题。就在他做胀缩实验的期间，这个希望居然有了曙光！很偶然地，他比较了一束绕射和入射的X射线强度，发现前者比想象中的弱很多，也就是说，大部分蛋白质轻原子的散射强度能由干涉而彼此相消。所以用一重原子（如汞）来取代它们，一定能产生可测得的强度改变。另一方面，如果用化学方法将重原子镶入血红蛋白中，似乎大部分的重原子试剂都会使蛋白质变性。

佩鲁茨1953年6月收到哈佛大学奥斯汀·瑞格斯（Austin Riggs）的一篇论文，使他更接近答案一步。瑞格斯研究镰形细胞贫血症时发现，血红蛋白中引入汞原子不影响其携氧能力。也就是说，加入汞原子并不改变血红蛋白结构!

佩鲁茨把血红蛋白晶体浸泡在汞溶液中，成功地在每个血红蛋白分子中掺入了2个汞原子，拍摄了血红蛋白晶体和掺入汞原子的晶体的X射线衍射底片。几个星期后，1953年夏天，当佩鲁茨提心吊胆地洗出了含汞血红蛋白晶体的第一张X射线照片时，两种血红蛋白晶体的衍射图案正有想象中的强度改变。佩鲁茨和布喇格在看两张照片时，意识到这个困扰他们16年的相位问题终于解决了。从这些差异，佩鲁茨能够能够计算出汞原子的位置，再从这些位置计算一套血红蛋白反射的相位。

佩鲁茨当时想，血红蛋白的结构最多花一年时间就能解决。令人失望的是，这次得意竟成了个幻影：因为晶体的“晶格混乱”，使得相位决定法无

用武之地。实际上，佩鲁茨花了 6 年时间才制作出第一幅血红蛋白的低分辨率结构，他一再证明“魔鬼总在细节中”：确定相位角需要数种重原子掺入的晶体［多重同型置换法，multiple isomorphous replacement（MIR）］，血红蛋白分子所有原子的坐标需要构建等，计算也有难度……

### 2.3.5 又是相位问题：核糖体结构的解析

30 多年之后，以色列女科学家阿达·约纳特（Ada Yonath）经过多年的孤独探索，终于制备出质量完美的核糖体晶体，衍射黑点图案很精细，足以推算核糖体三维结构。这时，相位问题又一次成为阻碍。

约纳特 1962 年在耶路撒冷希伯来大学获得化学理学学士学位，1964 年获生物化学硕士学位，1968 年，她获得魏茨曼科学研究所的晶体学博士学位。此后她先后在卡内基梅隆大学（1969 年）和麻省理工学院（1970 年）从事博士后研究，在那里开始接触蛋白质晶体学。1970 年她回到以色列，协助设立了以色列的首个也是随后十几年以色列唯一的蛋白质晶体学实验室。直到 1974 年约纳特才在以色列魏茨曼科学研究所稳定下来，独立开展结构化学研究工作。约纳特博士同时带领两个研究组：以色列魏茨曼科学研究所和德国汉堡马克斯-普朗克研究所的结构分子生物学中心。

约纳特是核糖体晶体学研究的先驱。她的研究开始于 20 世纪 70 年代后期，远早于大多数人认识到像核糖体这么大无规则结构的物质也能够结晶之前。她发展的晶体技术很快就得到广泛认可，成为常规实验技术。她巧妙地用嗜热和噬盐细菌的稳定核糖体于 1980 年制备到核糖体晶体。这些细菌的核糖体制备的晶体具有良好的 X 射线衍射性能，而大肠杆菌和其他来源的核糖体则未能成功结晶。约纳特还是冰冻结晶学（cryocrystallography）的先驱，那是一种快速冰冻晶体以最大程度降低强烈 X 射线对晶体的破坏的“闪冻”方法。约纳特也是第一个发现核糖体内部充满了孔道和小室，包括能够容纳翻译和新生肽的洞。这些发现一直遭到质疑，直到十多年后用冰冻电子显微镜进行了证实才结束。约纳特博士的创新技术和方法记录一直持续，在一篇论文中她描述了如何通过控制加热和 mRNA 类似物在核糖体晶体中启动蛋白质合成，再用各种化合物保存活化的核糖体晶体。后来她集中研究核糖体小亚基，在 2000 年 9 月份 Cell 杂志上发表了 3.3Å[1] 分辨率结构。

突破性发现往往来自先行者对未知领域的探索。这次的意志坚强的先行者是约纳特！20 世纪 70 年代末，她决定尝试用 X 射线测定核糖体的结构，这在当时被大多数人认为是不可能的事。实际上，蛋白质晶体学中 X 射线测定结构的第一步，也是最困难和最重要的步骤是制备蛋白质晶体，要求形

[1] $1Å=10^{-10}m$。

成的晶体几乎没有缺陷：分子以精确的方式重复排列。如果幸运的话，当盐水缓慢蒸发，就会生成美丽的晶体。但如果加热沸腾至干的话，只能得到一层白色粉末。条件不同，生成的晶体质量不同，X 射线晶体学即为这种情况。制备高品质蛋白质晶体是一件非常棘手的工作，蛋白质复合物越大，结晶的难度也越大，因此很多人怀疑约纳特想法成功的可能性。核糖体是最复杂的蛋白质和核糖核酸复合物之一。可以分为小亚基和大亚基两部分。人类核糖体小亚基包含 1 个长 RNA 分子和 32 个蛋白质分子；大亚基含 3 个 RNA 分子和约 46 个蛋白质分子。因此，每个亚基都包含数千核苷酸和氨基酸，换算成原子的话有数十万之多。约纳特计划建立包括核糖体中每个原子精确的空间位置的结构图。

约纳特决定制备核糖体晶体时，她选择从生活在极端环境下的细菌着手。*Geobacillus stearothermophilus* 能够在热泉环境下成活，热泉温度高达 75℃。约纳特的假设是要耐受这么高温度，核糖体必须极稳定，因此，其核糖体有可能形成品质高的晶体。1980 年，她成功建立了核糖体大亚基的第 1 个三维结构。尽管晶体并不完美，但也是一个巨大的成就。约纳特又投入了 20 多年时间，才建立含所有原子位置的核糖体结构图。

约纳特还进行了许多其他创新，如用液氮将晶体冷却到－196℃以稳定晶体，用其他适应性强的微生物的核糖体进行晶体实验，其中的一种与死海噬盐微生物 *Haloarcula marismortui* 非常相似。约纳特一步一个脚印，逐渐接近了目标。最终大家都认识到核糖体原子结构测定的可能性，越来越多的科学家加入了竞争队伍。其中包括托马斯·施泰茨（Thomas Arthur Steitz）和文卡特拉曼·拉马克里希南（Venkatraman Ramakrishnan）。施泰茨是美国生物化学家，1963 年在哈佛大学听佩鲁茨的报告受到启发，选择进入结构生物学领域。1966 年获得哈佛大学分子生物学和生物化学博士学位，曾在剑桥大学分子生物学实验室（LMB）作博士后，现为耶鲁大学分子生物物理学和生物化学系的斯特林教授，以及霍华德·休斯医学研究所的研究员。拉马克里希南 1971 年在印度巴罗达大学获物理学学士学位，1976 年在美国俄亥俄大学获物理学博士学位，1976～1978 年在加州大学圣迭哥分校获生物学研究生学位，1978～1982 年在耶鲁大学化学系做博士后。1999 年至今在英国剑桥大学 MRC 分子生物学实验室工作。

百万黑点图谱充满了意义。在 20 世纪 90 年代初，约纳特制备的晶体已经具有相当高的品质，图谱上的衍射点足够确定核糖体晶体中每个原子的位置。然而，还存在一个相当大的困难：又是 X 射线晶体学的相位问题！要推算结构，科学家还需要知道每个衍射点的相位角。这个数学信息与晶体中原子的位置相关。科学家确定相位的一个常用方法是把晶体用重原子（如汞）浸泡，重原子会附着在核糖体晶体的表面，通过比较有无重原子的晶体衍射

图，科学家就能够确定相角。而问题在于，核糖体太大了，附着的重原子很多，难以快速确定相位角，因此解决相位问题需要更多的信息。

托马斯·施泰茨（Thomas Arthur Steitz）成功解决了核糖体结构研究中的“相位问题”（phase problem）。施泰茨用电子显微镜专家 Joachim Frank 测定的核糖体图像，确定了晶体中每个核糖体的取向和位置（分辨率不足以确定每个原子的位置）。根据这些信息，加上重原子信息，最终成功推算出了相位角。1998 年施泰茨发表了核糖体大亚基的第一个晶体结构，就像一张模糊的照片，分辨率只有 9Å，看不到单个原子，但足以发现核糖体长 RNA 分子。这是一个决定性的突破，相位问题终于解决了！剩下的问题就是提高晶体质量，收集更多数据，以提高核糖体图像的清晰度。

从 1998 年初开始，有效数据不断开始积累。2000 年 8 月和 9 月，2009 年诺贝尔化学奖的三位获得者分别阐明了原子分辨率的核糖体晶体结构。施泰茨得到了 *Haloarculamarismortui* 核糖体大亚基的结构，约纳特和拉马克里希南得到了 *Thermus thermophilus* 核糖体小亚基的结构。这些成绩使在原子水平研究核糖体功能与结构的关系成为可能。

## 2.4 核磁共振方法

### 2.4.1 傅里叶变换和多维核磁共振

1991 年诺贝尔化学奖授予了瑞士苏黎世联邦理工学院的理查德·罗伯特·恩斯特（Richard R. Ernst），表彰他对核磁共振光谱（NMR）测定的方法学发展（methodological developments）。毋庸置疑，核磁共振是现代化学最重要的光谱学工具，能够用于研究溶液中小分子和大分子的精细结构，提供分子移动和相互作用的信息。术语“methodological developments”指新理论、新实验方法或革命性改进。简要回顾化学的发展，甚至整个自然科学的发展历史，不难看出，科学研究方法学的发展和进步常常对科学的进步具有重大影响。

恩斯特在科学生涯的初期就发展了傅里叶变换和脉冲核磁共振技术（Fourier transformation and pulse techniques in NMR spectroscopy），提高检测灵敏度数十倍乃至上百倍。核磁共振技术的“阿喀琉斯之踵”是灵敏度低，傅里叶变换的引入弥补了这个最大的缺点，从而可以进行少量物质中重要元素原子核的极弱信号的测定，如$^{13}C$ 和$^{15}N$。

核磁共振技术一个革命性的发展是恩斯特基于 Jean Jeener 的突破性思想发展的多维频率空间，二、三甚至更多维核磁共振技术，让科学家不仅能够记录化学位移，也能记录核旋之间的相互作用，分辨分子中特定原子被化

学键连接到哪个原子，以及周围空间存在哪些原子，化学反应中哪些原子发生了交换。

恩斯特回顾了核磁共振技术发展过程，他认为，核旋研究是理论物理学家也是实验物理学家的天堂，能够用最简单的实验装置检验量子力学和量子统计学的基本概念。核旋系统操作的简易性质也决定了其可以检验新实验概念。光谱学的通用方法主要在核磁共振领域得到重要发展，并在多个领域得到应用。实际上，核磁共振领域的重大实验进展不乏科学大师们的身影：伊西多尔·艾萨克·拉比（Isidor I. Rabi）建立了分子束磁性共振检测方法，为此他被授予 1944 年诺贝尔物理学奖；爱德华·米尔·斯珀塞尔（Edward M. Purcell）和瑞士物理学家费利克斯·布洛赫（Felix Bloch）发现了核磁共振精密测定物质结构和性能的方法，发展成为核磁共振成像（MRI 检查），这成为 20 世纪最重要的医疗设备之一，他们共享 1952 年诺贝尔物理奖；阿尔弗雷德·卡斯特勒（Alfred Kastler）与让·布罗塞尔（Jean Brossel）合作，他们研究了量子力学、光与原子之间的相互作用以及光谱，还研究了光共振和磁共振的结合，发展出光泵技术，提出了激光器和微波发射器理论。因为用于研究原子中赫兹共振的光学方法被授予 1966 年诺贝尔奖。此外，与核磁共振领域相关的诺贝尔物理奖获得者还有约翰·凡弗莱克（John H. Van Vleck），他发展了逆磁和顺磁理论，将矩量法引入了核磁共振领域；尼可拉斯·布隆伯根（Nicolaas Bloembergen）对早期弛豫理论和测定作出了重大贡献；卡尔·亚历山大·缪勒（Karl Alex Müller）对电子顺磁共振具有突出贡献；诺曼·拉姆齐（Norman F. Ramsey）则发展了化学位移和 $J$ 耦合（$J$ couplings）理论；汉斯·德默尔特（Hans G. Dehmelt）发展了纯核四极矩共振。

**矩量法**（moment method）

一种将连续的微分或积分方程离散化为代数方程的电磁场数值解法，广泛应用于积分方程的求解。它首先将待求场的微分或积分方程写成带有微分或积分算符的算子方程，再将待求场函数展开为某一组选用的基函数的线性组合代入方程，然后用一组选定的权函数分别对所得方程取内积（矩量），从而得到一个矩阵方程或代数方程组，进而解该方程得到展开基函数的各系数，从而得到待求场函数的近似解。它能够解决一些边界比较复杂的电磁问题，特别广泛应用于天线分析和电磁散射问题的求解。

核磁共振不仅激发了物理学家的广阔遐想，也吸引了越来越多的化学家、生物学家、医学家。核磁共振的魅力不仅在于概念的美妙，还在于它出乎意料的广泛用途。已发明出来一系列核磁共振工具，以增强其在多个领域

的功能，核旋系统的特性具有研究分子性质的先天优势。

自然的核探子位于被测量对象数飞米直径范围内的空间位置，能够感知最临近空间发生的变化，因此能够探测分子和物质的精细性质。这些探子与环境间的相互作用能量极低，小于 0.2J/mol，相当于 30mK 温度下热运动的能量，在检测分子性质时还几乎不存在扰动，相互作用对局部环境高度敏感，可通过核对子间相互作用采集几何信息，磁极子相互作用可提供距离信息，而 $J$ 耦合（scalar $J$ coupling interactions）又能进一步提供键角。

**理查德·罗伯特·恩斯特**（Richard R. Ernst）**小传**

理查德·罗伯特·恩斯特（Richard R. Ernst），出生于瑞士温特图尔（Winterthur，Switzerland），从苏黎世联邦技术学院于 1957 年获得化学证书，1962 年获物理化学博士学位。1963～1968 年，他在加州帕洛阿尔托（Palo Alto）从事化学研究，1966 年，与美国同事一道发现用强烈的短脉冲无线电波代替传统核磁共振光谱谱仪中使用的慢扫描电压波（sweeping radio wave）能够大幅提高核磁共振技术的灵敏度（当时只能测定有限几个原子核）。他的发现使测量微量物质和更多类型的原子核材料成为可能。1968 年，他回到瑞士母校任教。他对核磁共振领域的第二个重大贡献是发展出“两维”高分辨技术，从而能够研究此前的核磁共振光谱仪不能研究的“大分子”。恩斯特的创新成果让科学家能够测定生物系统的有机分子和无机分子，如蛋白质分子的结构，研究生物分子与其他物质如金属离子、水分子以及药物分子的相互作用，鉴定分子种类和研究化学反应的速度。为了表彰恩斯特在帕洛阿尔托瓦里安公司（Varian Associates，Palo Alto）发展起来的傅里叶变换核磁共振光谱仪和后来发展多维核磁共振技术这些杰出贡献，他被授予 1991 年诺贝尔化学奖。这些发现都巩固核磁共振在化学分析（核磁共振光谱）和医学临床检查（MRI 成像）的应用。1991 年他还获得路易莎总霍维茨奖（Louisa Gross Horwitz Prize）。

## 2.4.2 测定蛋白质结构的核磁共振方法

传统方法采用 X 射线晶体衍射方法测定蛋白质分子的空间结构。首先需要制备高品质蛋白质结晶，需要数月甚至数年时间收集 X 射线晶体衍射数据，再用计算机解析衍射数据建立蛋白质晶体结构。晶体中有规律、紧密地排列在一起的蛋白质分子被局限在晶格中，运动性较差，其空间结构与在生物细胞内蛋白质分子的本来结构可能存在较大的差别。而细胞中的天然蛋白质分子一直处于溶液状态，周围是水分子和其他的生物分子形成的微环境，

有一定的运动性。有些蛋白质只能稳定地存在于溶液状态，无法结晶。绝大多数种类的蛋白质也难以结晶，因此蛋白质结晶是X射线晶体衍射方法测定蛋白质分子的空间结构的瓶颈。

1978年，维特里希发现的核Overhauser效应（nuclear Overhauser effect）能够用于测量被键连接相邻原子网络的核磁共振测定蛋白质结构的方法。首先测定水溶液中的蛋白质样品，获得一系列不同的二维核磁共振图谱，然后从蛋白质分子的一级结构出发，通过比较和解析一系列二维核磁共振图谱，找到各个氨基酸各种氢原子在图谱上所对应的峰。再根据这些峰在核磁共振图谱上所呈现的相互之间的关系计算相应的氢原子之间的距离。在一级序列上相差甚远的两个氨基酸在空间距离上有可能很近，其氢原子的NMR信号峰之间就会呈现偶联和相互作用。关联信号的出现揭示两个氢原子之间距离小于0.5nm。蛋白质分子由几十个氨基酸残基组成的话就可以获得几千个与距离有关的信号，转换成对应的氢原子之间的距离；以这些距离参数和化学上有关原子与原子结合的参数，如原子间的化学键长、键角和原子半径等基本限制条件，用计算机程序模拟出该蛋白质分子的空间结构。这种结构测定方法避免了蛋白质结晶这个耗时费力的技术瓶颈，从而能更快、更可靠地确定生物大分子的组成和空间结构，是具有革命性的分析方法。在溶液中的蛋白质分子，更接近于其在生物细胞中的自然状态，因而在溶液状态测定的蛋白质结构也更真实。通过改变溶液的性质，能够模拟细胞内各细胞器中的各种微环境，更接近蛋白质发挥生物功能的结构。在溶液环境中，蛋白质分子结构表面的一些松散链段的运动性更容易表现出来，为研究蛋白质与蛋白质、蛋白质与底物或小分子的相互作用提供了一个有效的观察手段。

把核磁共振分析方法应用到了生物大分子（蛋白质和核酸）的结构研究中，建立了一套完整的确定生物大分子空间结构的革命性方法，这项开创性成就为维特里希赢得了1982年诺贝尔化学奖，成为该学院第二位因核磁共振方法而荣膺诺贝尔奖殊荣的教授。1991年恩斯特教授因其在核磁共振理论方面的贡献而获得当年的诺贝尔化学奖。

1976以后，维特里希、恩斯特、Kuniaki Nagayama、Anil Kumar共同发展出了多种多维核磁共振实验技术（2D NMR experiments），形成了核磁共振测定蛋白质结构的基础。1982年格哈特·瓦格纳（Gerhard Wagner）和库尔特·维特里希（Kurt Wüthrich）发表了碱性胰蛋白酶抑制剂（一个小分子量蛋白质）的序列特异NMR信号归属。同时，维特里希研究组的沃纳·布劳恩（Werner Braun）和蒂莫西·哈维尔（Timothy Havel）建立了从核磁共振数据计算蛋白质结构的算法和软件。1985年，迈克尔·威廉姆森（Michael Williamson）、哈维尔和维特里希报道了第一个蛋白质——精液

蛋白酶抑制剂ⅡA的溶液结构，但其研究结果一开始就遭到怀疑。直到有几个核磁共振方法解析的蛋白质结构被随后的X射线衍射方法解析验证后，核磁共振方法测定的蛋白质结构才开始被广泛认可。此后，核磁共振测定蛋白质结构技术又有了跨越性发展，设备的进步使能够利用的磁场越来越强。计算机处理器的进步使计算速度越来越快，基因重组技术可以将多种同位素富集掺入蛋白质样品中，使核磁共振测定技术从原来单一的核发展到三种甚至四种核在一个实验中同时共振而产生相关信号，由二维的核磁共振实验发展成三维甚至更多维的实验。核磁共振方法的应用也从蛋白质分子结构研究发展到研究蛋白质动力学、蛋白质与蛋白质、蛋白质与核酸及小分子的相互作用，以及药物筛选领域。

目前通过核磁共振测定的蛋白质的空间结构已经有几百种，占全部已测定蛋白质空间结构总数的20%左右。一部分蛋白质结构分别有核磁共振结构数据和X射线晶体衍射结构数据，结构对比揭示其骨架部分高度重叠，局部的表面区域呈现明显的环境诱导的差异。目前用核磁共振方法研究蛋白质的空间结构还局限于中、小蛋白质，即大约几百个以下氨基酸残基组成的蛋白质分子，而X射线晶体衍射的方法可以运用于真正的蛋白质分子。

## 2.5 蛋白质结构测定的其他辅助工具

### 2.5.1 实验方法

1962年，迈克尔·罗斯曼在剑桥大学发展了分子替代方法，能够从已知结构出发推测出相关蛋白未知结构的取向和位置。

1969年，班恩·修恩宝（Benn P. Schoenborn）证明中子衍射可以测定生物分子中氢原子的固定位置。

1983年，德国科学家贾桂·杜布守（Jacque Dubochel）用玻璃体冰冻方法成功制备出水化生物样品的非染色冰冻样品，玻璃化，或称之为生物样品无定形冰冻方法，已经成为冰冻电镜的关键技术。

1989年，哈肯·豪泼（Haken Hope）发展出在冰冻温度下采集生物大分子晶体衍射数据的新方法，极大降低了辐射对晶体样品的破坏。

### 2.5.2 计算及结构显示工具

1971年在美国布鲁克林国家实验室（Brookhaven National Laboratory）建立了蛋白质数据库（protein data bank），存放蛋白质和核酸的三维结构数据，开始只有7个结构数据，到2009年7月21日也只有59027个结构数据。目前由美国罗格斯大学（Rutgers University）的海伦·伯曼（Helen

Berman）主持维护。

1976年罗伯特·兰里奇（Robert Langridge）发展了第一个在计算机屏幕上显示蛋白质结构的程序“The nooks and crannies”，并在美国加州大学旧金山分校建立了计算机图像实验室。

1981年，珍尼·理查森（Jane Richardson）发展出用条带表示蛋白质结构的方法，简化了蛋白质结构展示的复杂性，增强了对蛋白质结构的理解，成为标准的理查森程序。

## 2.6 生物分子、复合物及病毒结构测定

### 2.6.1 重要生物学功能蛋白的测定

1965年，在伦敦工作的戴维·菲利普斯（David C. Phillips）测定了第一个酶的原子分辨率结构，该溶菌酶剪切糖链。

1967年，美国科学家弗雷德·理查兹（Fred Richards）、戴维·哈克尔（David Harker），理查德·迪克森（Richard Dickerson）、乔·克劳特（Joe Kraut）和威廉·利普斯科姆（William Lipscomb）开始发表解析的蛋白质结构，包括核糖核酸酶A和S（ribonucleases A and S）、羧肽酶（carboxypeptidase）、细胞色素c（cytochrome c）和胰凝乳蛋白酶（chymotrypsin）。

1968年，大卫·戴维斯（David Davies）、布赖恩·马修斯（Brian Matthews）等科学家开始解析抗体分子（即免疫球蛋白）的三维结构。

1973年，罗伯托·庖扎克（Roberto Poljak）发现免疫球蛋白折叠，一种抗体和其他分子结构中存在的结构模体。

1981年，唐·威利（Don Willey）测定了流感病毒的红细胞凝集素结构，有助于更深入了解病毒感染过程。

1985年，德国科学家约翰·戴森霍弗（Deisenhofer J.），罗伯特·休伯（Robert Huber）和哈特穆特·米歇尔（Hartmut Michel）等测定了紫细菌中光合作用反应中心、第一个膜蛋白结构，并因此三人分享了1988年度诺贝尔化学奖。

1989年，曼努埃尔·纳维亚（Manuel Navia）和亚历山大·乌鲁达瓦（Alexander Wlodawer）第一个解析出人类免疫缺陷病毒，即艾滋病病毒蛋白酶的晶体结构，该酶是病毒复制的关键之一。该结构的解析促进了治疗艾滋病药物的研发，对基于结构的药物设计作出了杰出贡献，使艾滋病这个对人类几乎致命的疾病转变为慢性病。

1998年罗德·麦金农（Rod Mackinnon）发表了第一个离子通道蛋白的

结构。离子通道蛋白是一类蛋白质，能够协助离子穿越细胞膜，行使神经冲动和其他重要生物过程的功能。这个结构以及随后的一系列结构极大拓宽了我们对神经科学的了解，促进了离子通道蛋白功能缺陷相关疾病（如囊性纤维化病、心律失常）的防治。麦金农因为这个杰出贡献获得了 2003 年诺贝尔化学奖。

同时，基因组研究的进展促进了基因组完整蛋白质结构的研究热情，标志着美国结构基因组研究的启动。

1999 年，以色列女科学家阿达·约纳特（Ada Yonath）、文卡特拉曼·拉马克里希南（Venki Ramakrishnan）、托马·施泰茨（Tom Steitz）和哈里·洛勒尔（Harry Noller）领导的研究组分别测定了核糖体结构，一个含多条 RNA 和蛋白质的从 RNA 碱基序列翻译成蛋白质氨基酸序列的分子机器。这是深入了解生命最基本过程——蛋白质合成机制的重要步骤，可能对新型抗生素的研发具有重要贡献。这些科学成就使约纳特、拉马克里希南和施泰茨分享了 2009 年诺贝尔化学奖。

2001 年，罗杰·科恩伯格（Roger Kornberg）研究组测定 RNA 聚合酶的三维结构。RNA 聚合酶能够催化从 DNA 转录合成信息 RNA、基因表达的关键步骤和控制点。这项杰出成就为科恩伯格赢得了 2006 年诺贝尔化学奖。

2007 年，布赖恩·科比尔卡（Brian Kobilka）研究组突破多种技术挑战，解析出人 G 蛋白受体（GPCR）的第一个三维结构。GPCR 不仅控制我们身体的许多重要功能，还和目前许多治疗性药物的作用机制相关。这些研究成就不仅能够加快更好疗效的新药开发，还能够加深我们对信号转导途径的了解。

2000 年 9 月，国立综合医学科学研究所（National Institute of General Medical Sciences，NIGMS）发起了为期 10 年的蛋白质结构计划（Protein Structure Initiative，PSI 或 PSI：Biology），目标为解析 10000 个蛋白质的三维结构。这些结构信息有助于揭示蛋白质在健康和疾病中的作用，指导基于结构的新药设计和研发。该计划一方面发展自动化蛋白质结晶技术，同时发展更快速的蛋白质晶体结构解析技术和使用更方便的软件；另一方面建设这些蛋白质结构数据共享的接口，供全球研究机构的科学家共享使用。2005 年，随着 10 个新的研究机构的加入，该计划进入第二阶段（PSI-2）。新建立了高通量结构生物学（PSI：High Throughput Structural Biology，PSI：HTSB），旨在通过集中组织研究人员在解决更大级别生物学难题方面进行高通量研究。项目涉及的 5 个主要部门分别是高通量结构检测中心、膜蛋白结构测试中心、高通量活性结构生物学研究联合部、PSI-SG（PSI Structural Genomics）知识库和 PSI-SG 资料档案库，形成 PSI 高通量结构生物学网

络核心。2009 年 2 月 15 日，NIGMS 宣布以促进运用已开发的技术研究生物医学重要研究难题为目标，出资促进生物学家和高通量结构检测中心合作的新方向。

### 2.6.2 病毒结构的测定

1964 年，又是在英国剑桥大学，阿龙·克卢格（Aaron Klug）证明 X 射线衍射测定蛋白质结构的原理也可以用于发展晶体电子显微技术，可以解析更复杂的结构，为他赢得了 1982 年诺贝尔化学奖。

生命有序，死亡失序。宇宙中自然发生化学变化的趋势是变得更加混乱无序。生命的特征之一是各层次结构的高度有序，从蛋白质和核酸分子到细胞器，到完整细胞，到组织器官，再到完整生物个体。生命有序的化学结构并不依赖于神秘的力量，而是建立在有序的分子结构和原子空间位置之上。在物质能够形成完美结晶的情况下，用 X 射线衍射能够测定分子中所有原子的空间位置。大分子倾向于聚集，生物功能也常常与分子聚集相关，如细胞核中染色体存在的形式——染色质，即是由 DNA 和上千种蛋白质形成的巨大分子聚集体。病毒是由核酸和蛋白质形成的较简单的聚集体。对病毒结构的研究能够为高等生物遗传物质结构提供线索。

大分子聚合体的研究难度很大，因此也很少能够制备晶体进行 X 射线衍射测定其结构。克卢格的科学贡献就在于发展出了测定生物系统中分子聚集体结构的方法。他的技术是电子显微镜与衍射方法原理的天才整合。电子显微镜早就用于对细胞亚结构进行成像，分辨力受到图像缺乏反差的局限。克卢格发现，即使看起来很平淡的图像也包含了大量的结构信息，使用数学工具能够挖掘出来。克卢格还用这种技术结合其他结构化学方法探索了病毒和染色质的结构。他对病毒的研究揭示了重要的生物化学原理，即细胞中分子的复杂聚集体是由其各个组成部分自发形成的！对染色质的研究能够揭示 DNA 上遗传信息阅读过程的结构控制机理；从更长远的观点来看，这些研究对阐明遗传物质如何失去对细胞分裂和生长的控制，即癌症的本质，具有极其重要的意义。

克卢格发展出的精妙晶体电子显微镜学（crystallographic electron microscopy）方法为测定宇宙中最精妙“活生物体”的化学结构提供了重要工具，并用于研究病毒、细胞核中蛋白质和核酸形成的染色质结构，澄清了重要生物化学机理。正因为这些重大基础研究贡献，他获得了 1982 年度诺贝尔化学奖。

此后，更多病毒高分辨率结构得到测定，如斯蒂芬·哈里森（Stephen Harrison，http：//crystal. harvard. edu/ ）测定的番茄丛矮病毒（tomato bushy stunt virus），迈克尔·罗斯曼（Michael Rossmann）随后测定了人们

更熟悉的病毒，如可引起感冒的鼻病毒（rhinovirus）的结构。

**番茄丛矮病毒病**

英文名称：tomato bushy stunt virus。

拉丁学名：*Tomato bushy stunt virus*，简称 ToBSV。

分类地位：+ssRNA 目、番茄丛矮病毒科、番茄丛矮病毒属。

形态特征：病毒粒体为球状，正二十面多角体，直径 30nm，分散在细胞质和液泡内，有晶体排列。ToBSV 不能被消化，即使通过人、畜的消化道也不能改变其传染性。体外存活期为几周。80～90℃加热 210min 可致死。

主要寄主：系统侵染番茄、曼陀罗等；局部侵染豇豆、菜豆、黄瓜、苋色藜等；也浸染矮牵牛、草莓、大车前、丹麦石竹、繁缕、金银花、锦葵、欧洲女贞、欧洲甜樱桃、苹果、葡萄、茄、甜瓜、郁金香、麝香石竹、香石竹等。

地理分布：阿根廷、加拿大、德国、法国、英国、意大利、摩洛哥、葡萄牙、突尼斯、美国。

危害部位：整株系统性侵染，产生枯斑。

侵染方式：必须在土壤积水的条件下借助汁液摩擦、病残体、种子、花粉、土壤积水等因子通过损伤的根细胞侵入。ToBSV 在河水中有所发现，因此用污水和淤泥作为肥料时，有可能传播该病毒。种子、花粉也能传毒。

危害症状：苗期染病 5 天后初显环状坏死斑，叶片褪绿且迅速落叶。该病在茎组织发展很快，苗期施用氮肥过多，造成茎变软，在靠近地面处出现坏死斑，引起猝倒。系统症状为幼叶卷曲或顶部坏死，由此引致侧芽增生或出现丛生。病株下部叶片褪绿和变紫。果实除产生褪绿斑以外，还在成熟果实上产生环状或平行斑，使番茄矮缩或产生斑驳。

http：//www. virology. ws/2009/03/06/virus-images-at-viperdb/

## 2.7 同步辐射光源及发展趋势

同步辐射（synchrotron radiation）是超高真空环境中以接近光速运动的电子在运动方向改变时产生的电磁辐射，其本质与我们日常接触的可见光和 X 射线一样，都是电磁辐射。1947 年，在同步加速器上发现这种辐射，因而被命名为同步辐射。1976 年，基斯·霍奇森（Keith Hodgson）等第一次证明同步辐射 X 射线衍射可以测定蛋白质单晶的结构数据，与传统 X 射

线衍射方法相比，具有更大的优势：

（1）强度高　是一般X射线光源的$10^4 \sim 10^7$倍，因此信噪比高，数据采集效率高，实验时间可大大缩短。如用X射线光机收集一个晶体衍射数据，通常需要7～15天，而利用同步辐射光源只需要十几秒或几分钟，工作效率提高了几万倍。

（2）能量可调性　同步辐射光的高亮度除了缩短了数据收集时间、提供高质量的数据外，更重要的是其能量可调性提供了结构鉴定所需的相位信息，使得结构鉴定时间从传统的数月至数年缩减到数周甚至数天。

（3）高度准直性和偏振性　能够获得很好的空间分辨率和微区内极高的光通量，可进行微束扫描，也可进行微量样品和微量元素的研究。使用第三代同步光源，蛋白晶体样品的尺寸可以减小到约20$\mu$m。

（4）脉冲性　同步辐射光是由与储存于环中周期运动的电子束团辐射发出的，具有纳秒至微秒级的时间脉冲结构。利用这种特性，可研究与时间有关的化学反应、物理激发过程、生物细胞的变化等。

此外，1990年，韦恩·韩椎柯森（Wayne Hendrickson）及同事们发展出硒代甲硫氨酸的解析蛋白质晶体结构的方法，这些硒原子在结构测定中能够发挥空间标志的作用。多波长反常衍射方法（multiwavelength anomalous diffraction，MAD），适用于在同步辐射装置上收集数据，是目前新蛋白结构的主要测定方法。因为波长反常散射法避免了结构优化中模型误差的直接引入和重原子与天然晶体之间的同晶性所产生的误差，可以方便快捷地解析出晶体的结构。MAD已经成为目前新型蛋白晶体结构测定的主要手段。

英国科学家J. Walker、美国科学家R. Mackinnon和R. Kornberg通过同步辐射研究生物分子的结构与功能，取得了突破性的成就，先后荣获1997年度、2003年度和2006年度诺贝尔化学奖。

在2003年我国出现SARS疫情后不久，我国科学家就利用同步辐射光成功测定了SARS病毒主蛋白酶的结构，为研制抵御SARS病毒的药物提供了重要信息。

目前世界上公认，同步辐射承担的分析量约占整个蛋白质结构的80%～85%。如果能够再提高同步辐射光束线的效能，则结构基因体研究的实现指日可待。

我国的同步辐射研究是从20世纪70年代末北京正负电子对撞机（BEPC）的建造开始的，起初是为高能物理研究而设计的，在1984年的一期工程期间决定一机两用，同时开展同步辐射的应用，这是第一代的同步辐射装置，称为北京同步辐射装置（BSRF，http：//www.ihep.ac.cn/），于90年代初建成。1983年中国科技大学获国家批准建设一台800MeV的低能第二代同步辐射源，于20世纪90年代初建成，称为国家同步辐射实验室

(NSRL，http：//www. nsrl. ustc. edu. cn/)，1992 年开始为用户服务。

上海同步辐射光源（简称“上海光源”；英文全名为 Shanghai Synchrotron Radiation Facility，缩写为 SSRF，http：//ssrf. sinap. ac. cn/）是我国迄今为止投资最大的国家重大科技基础设施建设项目之一，总投资约 12 亿元。上海光源是一台高性能的中能第三代同步辐射光源，能量位居世界第四，是目前世界上性能最好的中能光源之一。

在 SSRF 生物大分子晶体学光束线站上，可以利用以下多种实验方法开展生物大分子结构研究：多波长反常衍射方法（MAD），单波长反常衍射方法（SAD），同晶置换，分子置换等其他单波长实验方法。

基因体的 DNA 序列决定了所对应蛋白质内氨基酸的序列，而蛋白质在空间中的折叠决定了蛋白质的功能，所以想了解蛋白质的功能（也就是基因体的功能），必须先鉴定出蛋白质的三维空间分子结构。结构基因体学就是将一完整基因体所对应的所有蛋白质，鉴定出部分结构，并据以推测出所有对应蛋白质的功能。由于需要鉴定结构的蛋白质数量是如此庞大（以人类基因体为例，25000 个左右），依靠传统技术完成此研究的时间似乎遥遥无期，因为过去 50 年以来，全世界以传统技术鉴定的蛋白质结构总数仅约 1 万个。

2000 年 6 月 26 日，美国总统克林顿、英国首相布莱尔共同宣布，数国合作的“人类基因图谱计划”已完成，接着将进入后基因体时代。随即美国国家卫生研究院（NIH）所属之国家大众医药科学研究所（NIGMS）在 7 月 2 日宣布，在美国境内成立 7 个结构基因体中心，从事大量蛋白质结构鉴定工作。美国 NIH 随即成立六个任务中心，其中五个是同步辐射 X 射线设施，另一个是 NMR。基因的解码意味着新世代生命科学的来临，而结构生物学是开启此奥秘的钥匙。欧洲及日本亦相继跟进。蛋白质结构测定具有更辉煌的未来，发展高通量、高度复杂生物复合结构解析技术，建立预测和模拟蛋白质三维结构的计算机方法，普及各种不同蛋白质结构的生物功能和生物医学作用的探索，这些努力和发展无疑将增进我们关于生物学和健康领域的知识，有助于研发出更灵敏的疾病检测和诊断方法，研发治疗疾病的更高效药物，提高人类的生存品质。

## 参考文献


[1] Aue W P，Bartholdi E & Ernst R R. J Chem Phys，1976，64：2229-2246.

[2] Cherezov V，et al. Science，2007，318 (5854)：1258-1265.

[3] Cramer P，Bushnell D A & Kornberg R D. Science 2001，298 (5523)：1863-1876.

[4] Deisenhofer J，et al. J Mol Biol，1984，180 (2)：385-398.

[5] Dobson C M，Moore G R & Williams R J P. FEBS Lett，1975，51：60-65.

[6] Doyle D A，et al. Science，1998，280 (5360)：69-77.

[7] Electromagnetic radiation. In Encyclopaedia Britannica (2007). Retrieved May 26, 2007, from Encyclopaedia Brittanica online.
[8] Llerhand A, et al. J Am Chem Soc, 1971, 93: 544-546.
[9] Kumar A, Ernst R R & Wüthrich K. Biochem Biophys Res Commun, 1980, 95: 1-6.
[10] McDowall A W, et al. Journal of Microscopy, 1983, 131: 1-9.
[11] Miller M, et al. Science, 1989, 246: 1149-1152.
[12] Nagayama K, et al. Biochem Biophys Res Commun, 1977, 78: 99-105.
[13] Navia M A, et al. Nature, 1989, 33: 615-620.
[14] Perutz M. Nature, 1994, 371: 20.
[15] Philips J C, et al. Proceedings of the National Academy of Sciences, 1976, 73: 128-132.
[16] Rasmussen S G, et al. Nature, 2007, 450: 383-387.
[17] Richardson J S, et al. Nature Structure and Molecular Biology, 2000, 7: 624-625.
[18] Rossmann M G, et al. Nature, 1985, 317: 145-153.
[19] Rossmann M G, et al. Acta Crystallpgraphica, 1962, 15: 24-31.
[20] Schoenborn B P. Nature, 1969, 22: 143-146.
[21] Snyder G H, et al. Biochemistry, 1976, 15: 2275-2283.
[22] The Nobel Prize in Chemistry 2009 . Nobelprize. org. 28 Jun 2010 http: //nobelprize. org/nobel _ prizes/chemistry/laureates/2009/
[23] von Laue M. Nobel Lecture, November 12 (1915).
[24] Wagner G & Wuthrich K. Nature, 1978, 275: 247-248.
[25] Wagner G & Wüthrich K. J Mol Biol, 1982, 155: 347-366.
[26] Williamson M P, Havel T F & Wüthrich K. J Mol Biol, 1985, 182: 295-315.
[27] Wilson I A, Skehel J J & Willey D C. Nature, 1981, 289: 366-373.
[28] Wüthrich K & Wagner G. FEBS Lett, 1975, 50: 265-268.
[29] Yang W, et al. Science, 1990, 249: 1398-1405.

# 第3章 蛋白质结构基础理论

李荣秀

## 3.1 蛋白质结构形成的本质

生物分子之间专一和可再解离地形成复合物的能力是生命活动的基础。这种专一、可逆的相互作用被称为生物识别。每一种蛋白质分子的生物功能都毫无例外地依赖于与其他分子直接的物理接触。从基因的复制、转录和表达到基因的调控，从生物催化剂——酶的活性到激素的功能，以及有机体对外源有害物质（如毒素、病毒、细菌）的抵御和清除过程，无不依靠这种机制。细胞对接触到的每一种外源物质进行辨别和结合：细胞内的每一种物质都首先被合成或代谢它的酶识别和结合，细胞外的物质首先被细胞表面的受体结合，然后才能被吸收进入细胞。蛋白质与蛋白质能以非常专一和紧密结合的方式形成较大的复合物；蛋白质能与核酸结合，尤其在需要控制基因的复制与表达的情况下；蛋白质能与多糖，特别是细胞表面的多糖结合；蛋白质能与脂类结合，常常因此而整合到细胞膜上。

蛋白质-蛋白质相互作用几乎在细胞功能的每个水平都起着关键的运转作用：亚细胞器结构，跨膜的各种转运装置，染色质的包装，亚细胞膜微丝结构网络，肌肉的收缩和信号转导，以及基因表达的调控等。蛋白质-蛋白质相互作用异常会导致疾病，如早老性痴呆和进行性老年性痴呆。可以说，生物的结构、生长和繁殖都是以这种生物相互作用为基础的。普遍认为对分子相互作用基本机制的详尽研究是分子科学许多领域取得进展的重要基础。例如，在药物或分子传感器的设计方面，最首要的目标是发展一个可以与特定靶分子紧密并特异结合的化合物分子。在蛋白质由一级的氨基酸序列折叠成二级元件，再进一步折叠成三级结构过程中，蛋白质分子内部各基团间的相互作用势能启动相应的基团提供多肽链的摆动，各化学键的旋转找到相应的空间位置，使整个蛋白质分子结构的能量最低，从而形成相对稳定的蛋白

质高级结构。为了从理论上研究和模拟蛋白质结构，预测蛋白质分子低能量折叠结构，需要理解分子内部自由能的全景图谱。在酶分子的设计中，其方法是创造出一个可以与过渡态结构紧密结合的分子，同时能够避免与底物或产物分子结合得过度紧密。这些例子都建立在同样的概念框架上——原子和分子相互作用力和能量知识的应用。

## 3.2 分子势能

蛋白质结构是生物分子相互作用的物质基础。通常用数学方式表示生物分子的相互作用，但其本质比语言描述或数学公式复杂得多。虽然这样，我们仍然必须尽最大努力建立一套与事实相容的公式。习惯上，人们把原子间不同种类的相互作用称之为力（如范德华力），接着又给出一个特定相互作用的势能公式。

### 3.2.1 量子力学

量子力学是关于微观物质（分子、原子、电子等大小在 $10^{-10}$ m 数量级左右的物体）运动规律的理论。近代量子力学主要有四种形式，即：Schrodinger 的波动力学，Heisenberg、Born 和 Jordan 的矩阵力学，Dirac 的 $q$ 数理论和 Feynman 提出的途径积分方程。1927 年，Heitler 和 London 解析氢分子的量子力学方程，标志着量子力学原理开始应用于化学问题。量子力学方法中最通用的工具是分子轨道理论（MOT）。对于任意一个由原子核及电子组成的微粒体系，若忽略自旋轨道及其相关效应，则可用求解定态的 Schrodinger 方程：

$$H\Psi(R,r)=E\Psi(R,r) \tag{3.1}$$

式中，本征元 $\Psi$ 用来获取该体系的定态或散射态的结构信息。这里，定态的 Schrodinger 方程即为体系总能量 $E$ 算子的本征方程。$H$ 算子即 Hamilton 算子。$H$（Hamiltonian）运算符为：

$$\hat{H}=\sum_{i=1}^{n}\frac{\hbar^2}{2m}\nabla_i^2+\sum_{i=1}^{n}\sum_{j=i+1}^{n}\frac{z_i z_j e^2}{4\pi\varepsilon_0|\underline{x}_i-\underline{x}_j|} \tag{3.2}$$

式中，$m$ 代表粒子 $i$ 的质量；$z_i$ 代表粒子所带的电荷；$\hbar$ 代表约化普朗克常数。公式第一项表示经典的动力学能量；第二项表示静电能。所有分子，包括蛋白质分子的性质都可以用这个公式表示（除了相对论引起的微小效应）。然而，即使最简单的 $H_2^+$ 分子（含有两个质子和一个电子），这个方程式也无法解答。

### 3.2.2 Born-Oppenheimer 近似

对于由多粒子组成的分子体系，其定态的 Schrodinger 方程不易求解。根据原子的质量主要集中在原子核上，电子与核子在质量上有巨大差异，与核子质量相比，电子质量可以忽略。电子运动比核快得多，当电子运动时，可把核看成是固定的。Born 和 Oppenheimer (1927) 证明分子中电子和核子的分布可以近似地分开对待。对于一组核位置固定的系统，电子分布公式是成立的。可以用每一个原子核的位置来描述分子的结构，用 Born-Oppenheimer 近似键长或键角表达分子。由 Born-Oppenheimer 近似，体系的总波函数可以写为：

$$H\Psi_e(R,r)=E(R)\Psi_e(R,r) \tag{3.3}$$

式中，$E(R)$ 为体系中原子核坐标的函数，通常被称为 Born-Oppenheimer 面或位能面。通常把能够拟合这个 Born-Oppenheimer 位能面的表达式称为分子力场。选择合适的解析函数的形式及确定参数是确定一个分子力场的关键。量子力学方法分为从头计算法和半经验量子力学方法，常用的半经验量子力学方法有 AMI，MNDO 和 PM3 等。

根据近似，含 $N$ 个核子和 $n$ 个电子的分子的能量可用下列公式表示：

$$E_{\text{total}}=E_{\text{elec}}+\sum_{i=1}^{n}\sum_{j=i+1}^{n}\frac{z_i z_j e^2}{4\pi\varepsilon_0|\underline{R}_i-\underline{R}_j|} \tag{3.4}$$

式中，$R$ 代表核的位置矢量；$z_i$ 代表核子的电荷。公式第一项代表电子参与的相互作用对势能的贡献；第二项是典型的电荷作用项，这里表示分子中带电原子核间的排斥。电子波动方程和势能可由下列公式表示：

$$\hat{H}_{\text{elec}}\Psi(\underline{r}_1,\underline{r}_2,\cdots)=E_{\text{elec}}\Psi(\underline{r}_1,\underline{r}_2,\cdots) \tag{3.5}$$

运算符 Hamiltonian 中，电子对能量的贡献可用下列公式表示：

$$\hat{H}_{\text{elec}}=\sum_{i=1}^{n}\left\{\frac{-\hbar^2}{2m}\nabla_i^2-\sum_{j=1}^{n}\frac{z_i e^2}{4\pi\varepsilon_0|\underline{r}_i-\underline{R}_j|}\right\}+\sum_{i=1}^{n}\sum_{j=i+1}^{n}\frac{e^2}{4\pi\varepsilon_0|\underline{r}_i-\underline{r}_j|} \tag{3.6}$$

式中，$m$ 是电子的质量；$r_i$ 是电子的位置矢量。公式第一项的和相当于电子的动力学能量；第二项指的是电子和核子之间，以及和第三个电子之间的相互作用。

$$\hat{H}_{\text{elec}}=\sum_{i=1}^{n}\left\{\frac{-\hbar^2}{2m}\nabla_i^2-\sum_{j=1}^{n}\frac{z_i e^2}{4\pi\varepsilon_0|\underline{r}_i-\underline{R}_j|}\right\}+\sum_{i=1}^{n}\sum_{j=i+1}^{n}\frac{e^2}{4\pi\varepsilon_0|\underline{r}_i-\underline{r}_j|} \tag{3.7}$$

理论上，利用量子化学理论能处理所有的分子间相互作用。上述方程式的解答促进了量子化学领域的诞生与发展。量子化学方法的基本思想是找到

能描述分子性质，在原子核固定情况下的一组电子的分布（见 Born-Oppenheimer 近似法）。应用能量最小化原理，可使分子的几何结构最优化。从头计算量子化学从第一原理开始，不需利用外界或实验数据，允许考虑特殊种类，以及更常规的分子。可以跟踪化学反应，这些方法在有机化学中是非常有用的。

假设可以解出所研究的系统中所有原子核和电子（如酶、底物和它们周围的水分子环境）的薛定谔方程式。如果能在合理的短时间内完成方程求解，再与分子动力学（和其他技术）结合，就可能从第一原理出发，相当全面地处理一种蛋白质的结构性质、动力学和化学反应性质。不幸的是，应用这些方法解决蛋白质这样大的分子是不可能的，因为计算能力的局限只能解决几十个原子的分子。甚至用半经验量子力学，近似以及实验数据，也只限于考虑由几个氨基酸组成的小系统。然而，量子化学能够提供非常重要的参量，用于分子力学法和蛋白质中的化学反应研究。

对于最简单的分子，克服一定的困难也能够找出精确的数学答案。对于比较复杂的系统，有需要采用不同的途径。经常使用的 LCAO 方法，即把一个试验性分子轨道用整个分子中每个原子轨道的线性组合表示：

$$\Psi_{\mathrm{MO}} = \sum_k C_k \phi_k \tag{3.8}$$

差异原理（variation principle）可用于找出系数的数值。差异原理规定了计算出的 Hamiltonian 函数的期望值必须高于基础能态的值：

$$\langle E(\rho) \rangle = \frac{\int \rho^* \ \hat{H} \rho \mathrm{d}\tau}{\int \rho^* \rho \mathrm{d}\tau} \tag{3.9}$$

换句话说，如果我们有某个分子的试算波动方程，并作了一些降低能量的变化，此时，新方程比真正的波动方程能更好地近似于系统。因此，可以通过调整系数值获得 $\langle E(\Psi_{\mathrm{MO}}) \rangle$ 的最小值，从而找出系数的最佳值。

描述电子-电子排斥的 Hamiltonian 最后项使情况进一步复杂化。如果没有这一项，有可能用变量分离方法独立解决每一个电子的薛定谔方程，但是，当包括该项时，就不可能了。为避免这个问题，Hartree 和 Fock 提出并发展了标准近似法。假设由于核子和其他电子的存在，每一个电子都在平均场中运动，那么，即可猜测得到这个分子的最初波动方程。在 LCAO 方法中，选择出一系列被称为基础组的固定化 phi（至少必须代表每一个原子被占据的每一个原子轨道），最初猜测是一套系数（$C_K$）——（这必须符合 Pauli 原理，并且是反对称的）。在固定其他原子轨道的情况下，依次找出每一个原子轨道的最佳系数。继续这个过程，直到整个循环中所有的系数保持不变，并且自身相容为止。此方法被称为自身相容场法（self-consistent

field，SCF）。

这种方法通常被称为从头模仿法，它最大的优点是：处理一个分子时，不需要利用任何经验参数。如果基础数据的大小合理，并可降低 SCF 近似法引起的误差，通过在计算中考虑电子相关性，计算的准确性还是比较合理的。用能量最小化方法计算一系列核位置的能量（变化 $R_i$ 使势能最小），模型的几何性质能够得到改善。它的主要问题是计算费用，随 4.5 的 $n$ 次方升高，其中 $n$ 是用于计算的基础方程的数目。因此，提高小分子模拟和大分子结构的精度是非常昂贵的。实际计算的最高范围是几十个原子。半经验量子模拟法（例如 AMI）在 SCF 计算法中同时利用了近似和经验数据，能够扩大计算的范围。但是这种方法仍然太贵，甚至无法计算最小的蛋白质。

虽然计算费用昂贵，量子力学方法在蛋白质模拟研究中仍然非常有用，能够计算分子力学的参数，能够处理原子的部分电荷，修正分子在从头计算过程中的部分静电势。Bash 等（1987）采用半经验量子力学方法来处理系统中一小部分的原子（大约 20 个非氢原子），然后用 Warshel 和 Levitt 修改过的分子力学方法来处理剩下的原子。这种方法可以处理溶液中的化学反应，如磷酸丙糖异构酶的酶促反应。Warshel 等大量使用了一种化学键量子力学方法与分子力学方法的联合程序。另一种方法是以量子动力学的 Feynman 路径积分公式为基础的，已经用于常规分子力学力场，量子模拟亚铁细胞色素 c，模拟硬球流体和晶体中过量的电子。

模拟可以只在蛋白质的一小部分上进行。这允许处理不能用分子力学势能函数解释的过程，例如，化学反应或者光谱迁移。Arad 等用 AM1（一种半经验量子力学程序）模拟了木瓜蛋白酶反应机制中硫攻击的反应途径。根据分子力学势能方法对木瓜蛋白酶能量最小化，从得到的最低能量构象计算，该系统中酶活位点周围有 111 个原子。Penfield 等用氨代替组氨酸，甲基硫羟酸盐代替半胱氨酸，二甲基硫醚代替甲硫氨酸，7 个非氢原子对金属配位点进行近似；根据从头计算（*ab initio*）方法检测了质体蓝素中蓝色铜元素的电子结构以及铜元素光谱的超精细线分裂。Bajorath 等证明了酶（二氢叶酸还原酶）如何干扰结合中底物（叶酸）的电子分布。一种交互从头计算电子结构技术（局域密度功能理论，local density functional theory）能在库仑场中以量子力学方式处理原子数高达 49 个的底物。这个库仑场是由酶、辅酶和溶剂的每一个原子所带的部分电荷形成。

### 3.2.3 势能函数

已经简要回顾了蛋白质相互作用中几种重要的相互作用形式，并且能用简单合理的数学形式表达。这些相互作用力组合在一起形成势能函数公式。

$$\mathrm{PEF}(R)=\sum_{\text{bonds}} K_{\mathrm{b}}\{b(R)-b_{\mathrm{eq}}\}^2+\sum_{\text{angles}} K_\theta\{\theta(R)-\theta_{\mathrm{eq}}\}^2+$$
$$\sum_{\text{dihedrals}} \frac{K_\phi}{2}\{1+\cos[n\phi(R)-\gamma]\}+$$
$$\sum_{\substack{\text{non-bonded}\\ \text{atom pairs } i,j}} \left[\frac{A_{ij}}{r_{ij}(R)^{12}}-\frac{B_{ij}}{r_{ij}(R)^{6}}+\frac{q_iq_j}{\varepsilon_{\mathrm{r}}\varepsilon_0 r_{ij}(R)}\right] \tag{3.10}$$

此公式可用来计算一个具体蛋白质任何构象的势能［PEF(R)］值。构象可用坐标向量（$R$）表示。称其为势能函数，是因为它不包含原子运动的能量。可以用分子动力学方法计算这种能量。这个方程的目的是合理计算两个不同构象之间微观态能量的差别。这个能量的绝对值没有任何意义（肯定不是结构的自由能）。上述方程式也不适用于涉及化学键改变的任何过程，如模拟酶活位点的化学反应，也不能评价蛋白质是否采用某种折叠形式。用上述方程式计算蛋白质的势能时，还需要用到大量的参数（平衡键长，beq；键伸缩常数，Kb）。找出这些参数的过程是艰苦费力的，共同使用的势能方程只有大约四个（CHARMm，AMBER，GROMOS 和 ECEPP）。

虽然用势能方程（3.10）计算出的结果只是近似值，它最大的优点是计算费用低。这就允许表现真实环境，如直接模拟蛋白周围大量的水分子，也允许计算同一分子不同构象的势能值，促进了分子动力学方法的使用，如探测一个系统中的热运动。这些与量子化学方法形成鲜明的对比。用量子化学方法处理小系统也非常昂贵，以至于只能计算有限的次数，但可以获得非常准确的能量值。

采用量子力学方法的分子模拟软件有 Cerius2 INSIGHT Ⅱ（MSI），HyperChem（Hypercube）和 SYBYL（Tripods）。应该记住，许多的模拟软件都带有几种数学方法可供选择使用，例如，Mac/PC/UNIX SPARTAN（Wave function），SYBYL（Tripods）和 INSIGHT Ⅱ（MSI）结合使用分子力学、量子力学和（或）分子动力学。

## 3.3 分子内基团识别的要素

### 3.3.1 分子之间相互作用

真空中两个分子 $I$ 和 $J$，位于结合面上 $I$ 分子的所有原子和分子 $J$ 上所有原子相互作用时，两个分子相互作用的总能量 $E$，可以用 Clementi 公式计算：

$$E=\sum_{i}^{I}\sum_{j\neq i}^{J}\left[-\frac{A_{ij}^{\mathrm{ab}}}{r_{ij}^{6}}+\frac{B_{ij}^{\mathrm{ab}}}{r_{ij}^{12}}+\frac{C_{ij}^{\mathrm{ab}}\times q_i\times q_j}{r_{ij}}\right] \tag{3.11}$$

式中，总和指 $I$ 中的所有原子和 $J$ 中的所有原子；$r_{ij}$ 是原子 $i$ 和 $j$ 之间

的距离；常数 $A$，$B$，$C$ 是原子 $i$ 和 $j$ 相互作用的特征常数，取决于 $i$ 和 $j$ 的种类；上标 a 标示原子 $i$ 和 $j$ 的种类；上标 b 描述相互作用的原子类别（如氧原子 O，可以归为 OH、CO、CHO、COOH 等任何一类）；原子 $i$ 和 $j$ 的电荷用 $q_i$ 和 $q_j$ 表示。

公式第一项表示伦敦吸引力；第二项表示原子核之间的排斥力；第一和第二项的和描述几何互补性质；第三项描述库仑静电作用。Clementi 方程描述的是几何和构象固定不变的情况下两个分子的相互作用。分子项代表原子对的相互作用，只与原子的种类相关；而分母反映的是原子的位置，与原子的性质无关。所有的原子间力都能够归结为在原子核电场中运动的轨道电子的相互作用或轨道重组。因此，常常需要用量子理论进行分析，而用传统的力学则显得力不从心。生物识别研究中只考虑距离大于化学键，小于 0.35nm 之间的相互作用。尽管有一些经验处理方法，由于计算机计算能力的限制，还不能用量子理论处理所有的相互作用问题。

### 3.3.2 静电相互作用

带电原子的相互作用可以用库仑定律计算，但需要考虑到在一般情况下，电荷是离域化的，并表现出永久偶极（或多极）。在两种情况下，静电能量可以用部分电荷的概念或与化学键中心相关的非整数电荷计算，从而重现分子的静电场。

静电势代表所有永久和瞬时电荷分布，而与氢键形成相关的电荷需要单独处理。这类相互作用本质上是以下式表示的每对电荷的库仑相互作用：

$$E(\text{kcal/mol}) = \frac{332 \times q_1 \times q_2}{\varepsilon \times r_{12}} \tag{3.12}$$

式中，相互作用的电荷用 $q_1$ 和 $q_2$ 表示；电荷间的距离用 $r_{12}$ 表示，Å[1]；$\varepsilon$ 表示两个电荷间介质的介电常数。在水中，相互作用被水巨大的介电常数大大减弱，如：真空中两个正电荷间的氮原子间的相互排斥能量为 530kJ (130kcal)，在水中减弱到 33kJ，减弱倍数为 16（Fersht，1985）。蛋白质分子中的介电常数取决于电荷对的具体位置，如溶菌酶活性位点的有效介电常数大约为 5（Warshel & Levitt，1976）。这类相互作用的能量在 $-3 \sim -4$kcal/mol 之间，相当于在相应 pH 条件下氨基和羧基之间的相互作用大小。

按照相互作用的能量随距离减小的规律，电荷相互作用可以细分为三类：①离子对间的相互作用随距离的倒数（$1/r$）减小；②离子与电荷偶极子间的相互作用随距离的 4 次方的倒数（$1/r^4$）减小；③电荷偶极子间的相互作用随距离的 6 次方的倒数（$1/r^6$）减小。

[1] 1Å=0.1nm

### 3.3.3 氢键

氢键在配体与受体、底物与酶、蛋白质与 DNA 相互作用，以及蛋白质分子多肽链的折叠过程中都具有重要作用。

由于氢键在生物系统中具有非常重要的作用，它也许是研究得最多的非键相互作用。然而，尽管收集到关于氢键的事实数据不少，也利用了各种实验手段，包括测热法、光谱学、核磁共振（NMR）、X 射线和中子衍射，仍然没有可以普遍适用的氢键模型，形成氢键时能量的变化也是争论的焦点。氢键可以在质子供体基团 AH［其中 A 是电负性原子（O、N、S、F、Cl 和 Br）］和质子受体基团 B 之间形成。受体基团 B 具有一个孤电子对或不饱和键的 π-电子轨道，如氨基与芳香基团间的氢键（Waksman 等，1992）。氢键具有强烈的方向性，只允许 A-H-B 的角度在一定范围内变化。

因为氢原子上的部分电荷（约 0.28 电子电荷）是氢键相互作用的必要条件，非键作用、色散和排斥作用的分类也以酰胺氮原子为条件。这个模型得到 Allen（1975）提出的以从头计算分子轨道波动方程为基础的量子力学模型的验证。从蛋白质工程突变实验结果推测（Fersht，1987；Shirley 等，1992）出几种中性电荷——中性电荷氢键对自由能（$G_h$）的贡献为 −0.48～−1.9kcal/mol（−2～−8kJ/mol）。氢键 C—H···O═C 存在于晶体结构中。这种与电子云形成的氢键的自由能（$G_h$）大约为 −0.24～−0.48kcal/mol（−1～−2kJ/mol）。

### 3.3.4 溶剂或疏水相互作用

没有水，生命便不可能形成。在生物化学反应中，水不仅是一种良好的溶剂，也是在各种生化反应中提供 $H^+$ 和 $OH^-$ 的主要源泉。同时，水还是生物分子表现生理活性的唯一天然环境。水的结构特点及对生物分子作用力的影响可归纳如下：

（1）水分子由一个氧原子和两个氢原子组成，由于分子不对称的空间构型，其中氧原子具有负极作用，而相邻的两个氢原子则形成偶极分子中的正极。两个氢原子核与氧原子核连线之间形成的夹角约为 105°，是一个高度极化的极性分子，具有很高的介电常数。

（2）由于偶极分子中力场的作用，水分子彼此吸引而缔合，很少单独存在。每个水分子通过氢键可与另外 4 个水分子配价，形成四面体“冰晶”的结构，随着周围条件的变化，结构中的氢键的能力减弱，水分子还可以减弱离子键相互作用，而增强非极性疏水键的相互作用。

（3）通过氢键的缔合作用，许多生物分子如多糖、淀粉、蛋白质、核酸等分子中的羟基或氧、氮等原子都可以通过氢键与水结合，从而具有巨大的

水合或称“水化”能力。

通过上述关于一些生物分子中的结合力及分子间的作用力的简单介绍，可以看到生物体系中分子结构及分子之间联系的作用力十分复杂。作为一个生物分子，其基本骨架的各原子及基团之间都是共价结合的，整个分子的一级结构是比较稳定的。但分子与分子间的连接主要通过一些非共价键如氢键、电荷-电荷相互作用、范德华相互作用和疏水作用所维系，其相互作用的能量较弱，而且性质差别较大。

疏水键的概念出现在许多与溶剂作用相关的文章中，用于粗略描述存在于原子间、能够精确定义和可用方程描述的相互作用。建立理论模型的困难在于这个相互作用实质上与熵的关系比与焓的关系更密切，并且缺乏描述多个水分子行为的模型。

毫无疑问，一对蛋白质分子从相互独立到结合在一起经历了焓增加和熵减少的过程。从结合表面排斥开的溶剂分子经历了形成更多氢键更混乱的过程，从而导致焓减少和熵增加，因此，整个过程的自由能可表示为：

$$G=(H_{\text{protein}}+H_{\text{solvent}})-T(S_{\text{protein}}+S_{\text{solvent}}) \tag{3.13}$$

Nethethy & Scheraga 估计各种氨基酸侧链从水中转移到非极性环境中的 $S_{\text{protein}}+S_{\text{solvent}}$ 的数值在 $+9\sim+16$cal/(mol·℃)，取决于侧链的体积和性质。Chothia & Janin 更倾向于证明小溶剂（水）分子获得平移和旋转自由度时获得的熵值，$S$ 等于 $+70$cal/(mol·℃)，蛋白质侧链失去内部旋转自由度时失去熵值 $\Delta S$。两种模型预测的 $\Delta G$ 都大大低于零时，两个分子才倾向于结合。

习惯上，测量疏水作用的方式有两种：

第一种为微观方式，基于碳氢化合物从溶质表面转移到水中的能量；处于水中的特定碳氢链，Gibson & Scheraga 以及 Forsythe & Hopfinger 都认为，水化层中的每一个水分子被代替，其自由能变化为 $-130$cal/mol。对氢原子来说，$G$ 为 $+300$cal/mol；而对极性基团，如羰基氧原子、酰氨基氮原子，$G$ 为 600～4500cal/mol。Chothia 提出，水分子从蛋白质溶剂可及表面被去除时的自由能数值为 $-25$cal/Å$^2$。每一个水分子的横截面积大约为 6Å$^2$，因此每个水分子的自由能变化相当于 $G=-150$cal/mol。尽管 Chothia 模型可能是一种极端情况，早期溶剂可及表面的计算可能受到质疑，以上三种处理疏水键的方法所得到的自由能大体上一致。利用溶质分子的可及表面计算得到的疏水效果的数值在 20～30cal/(mol·Å$^2$)。正确评估摩尔转移自由能 $G$，可以用 Flory 和 Huggins 的混合理论计算，公式为：

$$\Delta G=-RT\left(\ln K_{\text{f}}+\left[\frac{V_{\text{s}}}{V_{\text{a}}}-\frac{V_{\text{s}}}{V_{\text{b}}}\right]\right) \tag{3.14}$$

式中，$K_{\text{f}}$ 是分配常数，用体积分数表示；$V_{\text{s}}$ 是溶质的摩尔体积；$V_{\text{a}}$

和 $V_b$ 是溶剂的摩尔体积。这个公式考虑到了溶质和溶剂分子大小的差别。用公式(3.14) 重新计算疏水作用得到的数值为 47cal/(mol·Å²)。通过拟合碳氢化合物的分子面积与从纯液体到水溶液中的转移自由能 [69.2cal/(mol·Å²)]，Tunon 等取得了与用碳氢化合物-水界面张力方法一致的结果 [69.2cal/(mol·Å²)]。

第二种，或称之为相互作用的宏观处理方式，来源于碳氢化合物与水界面的张力。已经总结的几种经验方法整合了分子疏水性、溶解性、溶剂分配系数等实验数据。Gibson & Scheraga，Hopfinger 以及 Forsythe 和 Hopfinger 采用了水化壳模型提供溶剂自由能一项。溶剂自由能是分子构象的函数，而分子构象是通过计算容纳溶质分子构象变化需要从第一水化壳去除溶剂分子的数目，最大可为整个水化壳中所有溶质分子的数目。对于每一个被替代的溶剂分子来说，自由能贡献需要考虑到溶剂和溶质中原子的性质。以类似的方式，Chothia 利用 Nozaki 和 Tanford 从氨基酸在有机溶剂中的溶解度计算出的疏水性数据以及 Richards 和 Lee 估计的可及表面数据 (Lee & Richards，1971) 得到了从蛋白质分子表面每移去 1Å² 的溶剂分子引起的自由能变化。另一种方法是用 Fersht 关于许多酶的数据，如 $CH_3$、$CH_2CH_3$ 和 $CH(CH_3)$ 的结合能数据，每个亚甲基减少 8kJ/mol。因为当一个亚甲基充填酶结构中的“孔”(不同于开放的结合部位) 时，得到结合能的变化，则每个亚甲基的疏水效果为 −2.5kJ/mol。当使用这个数值时，一个甲基可以算作 3 个 CH 基团。

### 3.3.5 伦敦类型的相互作用

非极性分子没有永久的多极性，但当与极性分子靠近时表现出诱导的或瞬时多极性效果。这些邻近 (永久或瞬时) 多极子或轨道电子运动的概率性质产生的扰动效应能改变分子中电子的分布。这些效果的强度取决于原子的极性常数，尽管每个原子看似很小 (与热运动相同)，但由于数目庞大，对分子内和分子间的能量贡献是相当可观的。理论上看，两个中性原子靠近时，一个原子中电子的任何扰动会诱导另一个原子中电荷的互补性变化；这种扰动引起的偶极子或多极子的相互作用能量是负数，即相互吸引。这种瞬时偶极子诱导现象首先由 London (1937) 提出，因此称为伦敦色散力 (London dispersion forces)。如上所述，这种量子理论预测的分散力起源于电子分布的瞬时扰动，即使在非极性分子间也存在着这种色散力，频率在光波范围内 ($10^{15}$ Hz)。实际上，相互作用是在瞬时偶极子和诱导偶极子之间发生的。因此，色散力在形式上与诱导作用类似，但强度取决于相互作用的基团的极化能力。

瞬时偶极子与诱导偶极子之间的相互作用在形式上与诱导相互作用非常类似，但相互作用的大小由相互作用的两个基团的极化能力决定。这种相互

作用起源于量子力学性质，因此即使对于简单的分子，也需要量子表达方式才能与试验数据一致。Slater 和 Kirkwood（1931）提出的公式如下：

$$v=-\frac{3\pi\times e_o^2\times a_o^{1/2}}{\varepsilon^2 r_{12}^6}\times\frac{\alpha_1\alpha_2}{(\alpha_1/n_1)^{1/2}+(\alpha_2/n_2)^{1/2}} \tag{3.15}$$

式中，$e_o$代表电荷；$a_o$ 为波尔直径；$r_{12}$是距离，Å；$\alpha_1$ 和 $\alpha_2$ 是基团的极化常数，Å$^3$；$n_1$ 和 $n_2$ 是相互作用体系的外层电子数，$\varepsilon$ 是有效介电常数。London（1937）对这种作用进行了更精确的处理：

$$v=-\frac{3}{4}\times\frac{I_1 I_2}{I_1+I_2}\times\frac{\alpha_1\alpha_2}{\varepsilon^2 r_{12}^6} \tag{3.16}$$

式中，$I_1$ 和 $I_2$ 是相互作用基团的离子化能，kcal/mol。

应该认识到相互作用的物理基础是以光的频率进行波动的结果，两个公式需要用高频极限下的介电常数。$\varepsilon$ 的数值相当于溶剂折射率的平方，即使在像水这样的极性溶剂中也是这样。与此形成鲜明对比的是在静电相互作用中，$\varepsilon$ 的数值相当于零频率下的介电常数。

### 3.3.6 原子核在超短距离时的排斥作用

其他非键作用是在短距离时产生的排斥作用（$r_{12}<2\sim3$Å），而且随 $r_{12}$ 趋近于 0Å 而趋于无穷大，因此与原子和分子具有空间填充性质一致。原子间距离对核排斥作用的影响非常强烈，但对于函数的形式还没有达成一致。描述核排斥作用的方程之一是 Lennard-Jones 方程：

$$v_r=\frac{+B}{r_{12}^{12}} \tag{3.17}$$

式中，$r_{12}$是一对原子之间的距离；$B$ 是一个只与原子对相关的常数。从这个方程式可以明确看出，分子识别过程中任何不互补的表面都会引起巨大的排斥，并进一步造成分子复合物的解离。

## 3.4 识别的专一性

在分子结合的过程中，通常是溶液中的两个分子相互靠近形成某些构型的复合物（新分子）。参加反应的分子通常被称为配基和受体。实际上，从技术的观点看，怎样命名并不重要。总之，由于反应物和复合物的构象自由度，每一对反应物可能会以多种构象相互结合，形成多种复合物分子或者同一种复合物的多种构象。

人们对分子科学研究的兴趣大多集中在分子或者分子片段的结合作用方面。伴随着结合过程产生结合自由能。由于分子的相互结合，溶液中总自由能发生变化，与此同时复合分子的浓度也发生变化。自由能的变化量决定了

结合的分子复合物在平衡时的比例。在很多情况下，总的结合自由能可被分解成独立的静电和非静电作用；静电结合对自由能的作用完全是由于反应物所带的静电荷引起的。天然分子结合反应的表面驱动力大部分是非静电作用，包括所谓的疏水作用。对疏水作用有利于分子结合的解释之一是：水分子之间的相互作用比水分子与绝大多数的溶质分子相互作用要有利得多。结合反应减少了溶质-溶剂的净分界面积。需要更少的溶剂分子包围溶质，因而溶液的自由能减少，导致结合的倾向。然而，按照同样的标准，疏水相互作用没有提供结合特异性的机制，例如，选择性结合一个分子而不是另一个具有相似形状的分子。通过使分子表面形状互补可获得一些结合的特异性，如果配合得好，范德华相互作用大。近来的大多数计算表明，静电相互作用实际上不利于结合，静电相互作用对结合自由能的贡献几乎总是正的。与不带电荷的分子相比，使分子的结合变得更难。在天然分子复合物中，静电相互作用通常被用来创造特异性。形状相似，静电分布不同的反应物，产生明显不同的结合自由能。

静电相互作用对结合的贡献明显是非直觉的，因为相关的结合经常发生在水溶剂中，通常牺牲反应物和水之间有利的静电相互作用支持反应物之间的有利（通常是不很有利的）相互作用。分子结合研究中出现的一个问题是：静电相互作用对分子间结合的贡献是否是不利的？

任何一种相互作用对分子识别的贡献都无法精密确定。主要原因是分子间距离、相互作用的相对几何位置和形状，以及局部溶剂环境的难以确定性。一般来说，在最佳距离时，相互吸引的电荷-电荷相互作用的结合能大约为－5kcal/mol，电荷-偶极子相互作用的结合能大约为－4kcal/mol，偶极子-偶极子相互作用的结合能大约为－3kcal/mol，四极子的弱相互作用的结合能大约在－1～－2.5kcal/mol 之间。疏水相互作用的能量约为－69.2cal/(Å$^2$ · mol)。Andrews 等研究认为在考虑到溶剂分子贡献的基础上，离子键、氢键和范德华力对自由能的贡献分别在－10kcal/mol、－3kcal/mol 和－1kcal/mol，而在结合时每失去一个内部自由度的自由能为＋0.7kcal/mol（表 3.1）

**表 3.1　非键作用能量的经验值**

| 相互作用的类型 | | 结合能/(kcal/mol) |
|---|---|---|
| 离子键 | | －3.0～－4.0 |
| 氢键 | | －0.48～1.9 |
| 疏水相互作用/(1/Å$^2$) | 分子表面 | －0.069 |
| | 溶剂可及表面 | －0.020～－0.030 |
| 范德华相互作用 | | －1.0 |

这些力对结合专一性的影响是不同的。离子键和氢键的专一性最强，疏水作用和伦敦色散作用最不具有专一性，如表 3.2。实际上，不会仅因一种相互作用而发生生物识别，它是所有这些相互作用的综合。

**表 3.2　化学基团间的相互作用**

| 吸引作用 | 对结合能和专一性的贡献 |
| --- | --- |
| 专一性<br>和<br>结合能贡献<br>↓<br>减少 | 离子键<br>氢键<br>电荷-永久偶极子<br>电荷-诱导偶极子<br>多极子-多极子<br>多极子-诱导多极子<br>范德华相互作用<br>疏水作用 |
| 排斥作用 | 对专一性的贡献 |
| 专一性 | 空间位阻<br>静电排斥作用 |

## 3.5　基团间识别与接触

分子识别和结合过程是两个分子形成复合物时获得自由能的结果。这种形成复合物的过程可以用范德华作用、静电作用、氢键和疏水作用描述。多亚基蛋白，如血红蛋白和蛋白-蛋白复合物（如胰蛋白酶和胰蛋白酶抑制剂）的内部界面是致密充填和互补的 。所有的氢键供体和氢键受体是相互配对的，所有正电荷和负电荷也相互配对。两个表面在几何上相互补充。互补有两种性质：①立体形状互补，因为两个分子结合时不能互补的表面由于原子间的距离太近造成巨大的排斥作用；②吸引相互作用的互补，包括电荷互补（相互结合的表面必须具有相反的电荷性质，从而形成吸引作用而不是排斥作用）、疏水表面互补（疏水和亲水表面倾向于与性质相同的表面相互作用，从而使溶剂的效应最大）、氢键供体和受体互补（使氢键的形成成为可能）。

有许多蛋白质与配体和其他分子相互作用的实例（表 3.3），其中自由能的差异是结合专一性的基础。由于两个相互作用表面的立体空间必须互补，电荷和氢键相互作用要求严格的相互作用对象，任何的错配都会减少结合能。Chothia & Janin 认为作用的专一性是由氢键的形成决定；Fersht 用点突变研究酪氨酰-tRNA 合成酶的结果显示，改变不带电荷的基团使氢键形成受到影响时，对专一性的改变是 1.5～2.5 倍，改变不带电荷的基团引起空间排斥时，对专一性的影响是 1000 倍。Popplewell 等利用蛋白 A 和抗体 IgG 作为模式结合系统研究未配对的氢键受体和供体对两个分子的结合能量的影

响为 1.3kJ（0.31kcal）。

**表 3.3　化学基团间相互作用**

| 性质 | 受体 | 配体 | 相互作用 | 距离 |
|---|---|---|---|---|
| 形貌 | 凹(凸) | 凸（凹） | 诱导 | 短距离($r^6$) |
| 表面电势 | 正(负) | 负（正） | 电荷 | 长距离($r^1$) |
| 氢键 | 受(配)体 | 配(受)体 | 偶极 | 短距离($r^3$) |
| 溶解性 | 非极性(极性) | 极性(非极性) | 诱导偶极 | 短距离($r^6$) |

尽管范德华相互作用和疏水相互作用不需要严格的配对，它们也从其他方面对两个分子的结合有贡献。Chothia & Janin 总结为：疏水作用是两个分子形成复合物的稳定力量，也证实了 Kauzmann 的假说，即熵项稳定蛋白的聚合，小分子的结合主要是疏水相互作用的结果。在极其稀少的情况下，配体与结合部位的亲和性由一种相互作用决定，如某些抗原-抗体复合物（van Oss 等，1982）。数种相互作用对一个亲和反应都有贡献是正常的，尽管每种相互作用的相对贡献从一个复合物到另一个复合物会有所变化。

## 参考文献


[1] Arad D, et al. J Am Chem Soc, 1990, 112: 491-502.

[2] Baum J O, et al. In "Supercomputational Science." (Evans, R. G. & Wilson, S. eds.). New York: Plenum Press. 275-286 (1990).

[3] Karplus M, et al. Faraday Discuss, 1992, 93: 239-248.

[4] Penfield K W, et al. J Am Chem Soc, 1985, 107: 4519-4529.

[5] Singh U C & Kollman, P A. J Comp Chem, 1984, 5: 129-145.

[6] Warshel A, Levitt M. J Mol Biol, 1976, 103: 227-249.

[7] Zheng C, et al. Nature, 1988, 334: 726-728.

# 第4章 从序列到结构

连鹏　陈仕高　魏冬青

## 4.1 背景介绍

随着实验技术的进步，蛋白质测序日趋成为一种常规手段。从 NCBI RefSeq 数据库（http：//www. ncbi. nlm. nih. gov/refseq）的数据来看，仅从 2010 年 5 月 9 日到 2010 年 7 月 13 日新测序并被注释的序列数量就从 3549618 条增加到了 3641224 条，平均大约每分钟就有一条新的序列被确定！这是序列库，蛋白质结构数据的情况怎么样呢？截至 2010 年 8 月 31 日 17 时，PDB 数据库（http：//www. pdb. org）中已测结构的蛋白质为 67656 个。与序列库相比，这个数据量简直少得可怜！需要说明的是，这其中还有相当数量的冗余。然而，从应用的角度来讲，无论是医学上的药物设计还是生物技术中的酶设计，需求更多的都是蛋白质的结构信息。基于结构的药物设计、酶设计等针对性强，方便、快捷、可靠，并且已经得到了相当广泛的应用。如果仅仅针对已测结构的蛋白质进行设计改造的话，这 6 万多的数据量是远远不够的。从另一方面讲，序列数据的庞大也为结构研究提供了一种潜在的资源。根据当今生物学家的普遍观点，蛋白质的一维序列决定了蛋白质的高级结构。也就是说，仅从序列出发，依照某种规律是可以将蛋白质折叠成特定的高级结构的。如果有一天这种规律被完全揭示的话，那么就能够极大地丰富蛋白质的结构信息，将相关领域的研究提高到一个更高的、全新的水平。然而蛋白质结构预测，依据已知的序列来预测蛋白质的高级结构，这是困扰生物学家、化学家、物理学家、数学家等多年的古老话题。在过去的四五十年里人们进行了各种各样的尝试，并取得了一些成果，但是到目前为止仍然没有哪种方法能够仅从序列出发得到蛋白质的合理的高级结构。

那么蛋白质的高级结构到底是什么样的呢，为什么这么难以预测呢？我们知道，蛋白质是由单条或者多条氨基酸链弯曲折叠形成的。它的结构从简

单到复杂可以分为一级结构、二级结构、三级结构和四级结构。一级结构就是组成蛋白质的氨基酸序列，序列通过不同的方式弯曲可以产生 α 螺旋、β 折叠和无规则卷曲等二级结构，二级结构经过折叠、组合可以形成特定的三级结构。一般而言，由三级结构组成的蛋白质结构域就有一些简单的功能了，但是要完成生物体内复杂的化学反应还需要将多个结构域按照一定的方式组合起来，这就是蛋白质的四级结构。

蛋白质结构预测依据所预测结构的不同可以分为二级结构预测、三级结构预测和四级结构预测三个大的类别。下面我们针对这三个层次的预测分别作一简单介绍。

## 4.2 蛋白质二级结构预测

二级结构预测的主要目的就是预测蛋白质的序列可能形成的二级结构，哪些地方形成螺旋、哪些形成折叠以及哪些形成卷曲。这种预测从 20 世纪 70 年代的 Chou-Fasman 方法经过 GOR 方法，到今天的机器学习方法，预测准确性从 60%提高到 80%以上，接近二级结构预测的理论上限 90%。

### 4.2.1 Chou-Fasman 方法

Chou-Fasman 方法是早期依据经验来预测蛋白质二级结构的一种方法。这种方法是在部分蛋白质结构已经被解析之后出现的。经过对已解析蛋白质中氨基酸分布的统计分析，人们发现，氨基酸在蛋白质不同二级结构中的出现是具有一定的偏好性的。比如，丙氨酸、谷氨酸、甲硫氨酸等有很高的概率出现在螺旋的起始部分；而甘氨酸和脯氨酸则大多出现在螺旋的末端。依据这些偏好性就可以预测序列的二级结构了。这种方法简单明了，速度也很快，但是预测的准确性太低。初期可能是因为蛋白质晶体结构少，这些氨基酸偏好性参数较粗糙，但是最近人们已经已知的晶体结构更新了氨基酸的参数，这种方法所能达到的准确率也才只有 50%～60%。这是因为决定蛋白质二级结构的因素很多，氨基酸的种类是一个很重要的因素，但并不完全是决定性的。氨基酸与周围氨基酸的相互作用以及所处结构环境对于决定该氨基酸的二级结构同样有着不可忽视的作用，而这种方法忽略掉的恰好就是这些因素。尽管这种方法的预测准确率不高，但是它对蛋白质二级结构预测进行了早期的尝试，并为日后的预测提供了一种可行的思路。

### 4.2.2 GOR 方法

GOR 方法（Garnier-Osguthorpe-Robson 方法）是继 Chou-Fasman 方法出现后不久发展出来的另一种基于氨基酸概率参数的方法。与 Chou-Fas-

man 方法不同，这种方法除了加入氨基酸在不同二级结构中出现概率的参数外还引入了一种限制概率，来衡量该氨基酸周围已经存在的二级结构对它的影响。另外，这种方法采用贝叶斯统计模型来进行预测。GOR 方法对螺旋和折叠的预测有了较为显著的提高，但是对转角和无规则卷曲的预测还存在明显的不足。这种方法总的预测准确率在 65%左右。近年来，经过 Garnier 等人的不断发展和完善，这种方法的准确率可以提高到 73.5%，并且支持在线预测。他们的网址为：http：//gor. bb. iastate. edu。

### 4.2.3 机器学习

生命科学的发展为物理、化学、数学、计算机等学科提供了丰富的研究对象，机器学习方法在蛋白质二级结构中的应用就是一个很好的例子。现在使用最多、准确率最高的应该就是机器学习方法了。机器学习方法主要有两种：人工神经网络方法（artificial neural network，ANN）和支持向量机方法（support vector machines，SVM）。

人工神经网络也称为神经网络，是受生物神经网络工作原理的启发而建立起来的一种数学模型（图 4.1）。它是由大量节点（神经元）相互连接组成的运算模型。神经网络中每个神经元代表了一种特定的激励函数，神经元之间连接是这两个神经元之间的信息流通量的权重。神经网络的输出值随着网络中神经元激励函数和神经元之间权重的不同而变化。用已知的信息优化神经网络中各神经元的激励函数和权重的过程就是神经网络学习的过程。这个过程也称为训练，训练好的神经网络可以用于预测。同样地，蛋白质二级结构预测中，可以先用已知晶体结构中氨基酸序列和所处二级结构的信息训练神经网络，然后再用来预测其他序列的结构。尽管不能有效地预测远程氨基酸之间的氢键相互作用，但是这种方法还是相当可靠的，其预测准确率可以达到 70%以上。最近，英国邓迪大学（University of Dundee）Barton 课题组发

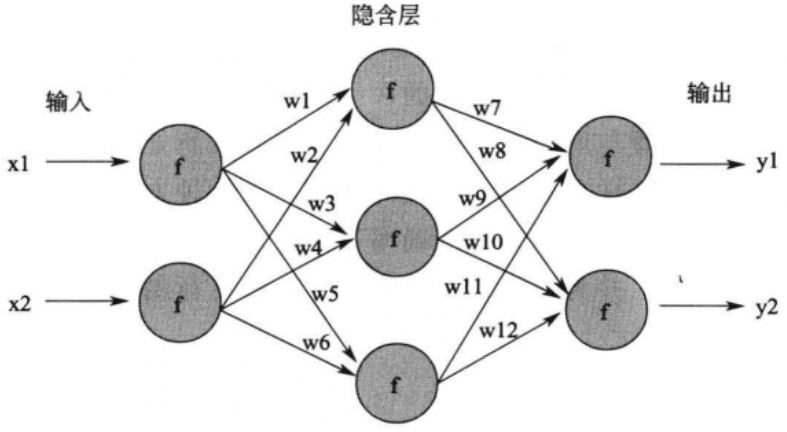



图 4.1 人工神经网络示意图

展的基于神经网络的预测程序已经可以将预测准确率提高到81.5%，他们的在线预测网址为：http：//www. compbio. dundee. ac. uk/www-jpred。

支持向量机是采用监督式学习方法的，它是在统计分类和回归分析中广泛应用的一种分类器。这种分类器本质上是一种线性分类方法。由于线性分类方法简便易行，在很多领域得到了广泛的应用。但是对稍微复杂的问题，线性方法的弊端也是显而易见的。例如，图4.2(a) 中的分类问题，很难用一条直线将圆形和方形两种变量分开。遇到这种问题，线性分类方法是不是就完全失灵了呢？答案是否定的。SVM方法提供了一种新的思路，是的，在二维平面里面不能找到一条直线将圆形和方形的变量分开，那么如果我们把这两个变量映射到更高维度的空间，比如三维空间中，会怎么样呢？如图4.2(b) 所示，在更高维度的空间中很容易找到一个或者一组超平面（hyperplane）将这两个变量分开。这就是SVM方法实现复杂问题分类的方法思想。由于SVM方法具有较好的分类预测能力，因此这种方法也被用于蛋白质二级结构和蛋白质骨架二面角预测中，并且取得了相当好的结果。同神经网络方法一样，SVM方法也需要先用已知的结构信息训练，然后才能预测，因此也属于机器学习的方法。YASSPP是一个利用SVM方法进行预测的服务器，其准确率可以达到79.3%，网址为：http：//glaros. dtc. umn. edu/yasspp。

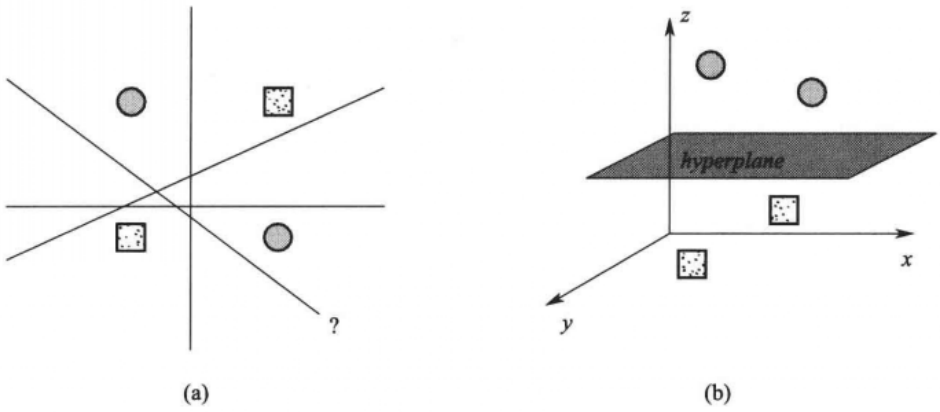



图4.2　SVM分类思想示意图

如今，机器学习方法预测蛋白质二级结构已经得到广泛应用，通过选取适当的训练集和调整结构、参数，能够将预测的准确率提升到80%以上。

## 4.3　蛋白质三级结构预测

尽管蛋白质三级结构预测比二级结构预测更具现实意义和实用价值，但是与二级结构的预测相比，蛋白质三级结构的预测要困难得多。到目前为止，三级结构预测仍然是一个极具挑战性的课题。它所面临的主要问题是蛋

白质自由能的计算和蛋白质自由能全局最小值的搜索。一个好的预测要遍历蛋白质所有可能的构象，而蛋白质构象是由氨基酸所有可能构象排列组合得来的，因此蛋白质构象本身就是一个天文数字。要从蛋白质所有构象中找出能量最小的构象，困难程度可想而知。一种相对可行的办法是同源建模。在目标序列的所有同源蛋白质中，如果某一个或者几个的晶体结构已经被解析，并且它们与目标序列的相似性在40%以上，那么就可以以同源蛋白质的晶体结构为模板来预测目标序列的三维结构。这种方法计算量小，并且预测结果比较可靠，因此得到广泛应用。

### 4.3.1 蛋白质结构的从头预测

蛋白质结构的从头预测是仅从序列出发，依据物理学和蛋白质折叠基本原理来进行三维结构预测的一种方法。与其他方法不同的是，它不依赖任何已经解析的蛋白质结构，仅根据理论进行预测。它对物理学、化学、结构生物学和结构基因组学都具有潜在的重大意义。但是这种方法的预测是非常困难的，它所面临的问题主要是如何对巨大的蛋白质构象空间进行搜索和如何寻找蛋白质自由能的全局最小值。目前，一种可行的解决办法是先将20种氨基酸所有可能的构象存储起来做成“候选构象”或者“诱饵构象”（candidate structure or decoy structure），然后依据打分函数（scoring function）来引导构象空间的搜索。氨基酸每变换一种构象就依据打分函数中的规则对体系打分，然后根据得分的不同将构象排序优选出最佳构象。常用的打分函数有两种：一种是基于数学模型和物理学原理的打分函数；另一种是基于统计模型的打分函数。这种方法中打分函数起了非常重要作用，因此选择一个合适的打分函数是这种方法的关键。

蛋白质结构的从头预测是一个庞大的工程，它不仅需要好的算法，还需要巨大的计算资源。目前，除超级计算机如IBM的Blue Gene和日本RIKEN研究所的MDGRAPE-3以外，一些分布式计算项目如Folding@home、Human Proteome Folding Project和Rosetta@home等也在尝试解决这一难题。

### 4.3.2 蛋白质结构的同源建模

同源建模是目前从序列预测结构的最切实可行、应用最广泛的一种方法。上海交通大学生命科学与技术学院生物信息学实验室魏冬青教授课题组就运用这种方法对细胞色素酶CYP450家族的许多成员进行了同源建模，并以此为基础开展了一系列的酶催化机理、药物设计等相关的研究工作。同源建模方法使用与目标序列同源的某一蛋白质的实验结构作为模板，对目标序列进行三维结构的预测。使用这种方法有两个前提条件：首先，目标序列的

同源蛋白质中某一个或者几个结构已经被解析（X 射线或者 NMR 结构均可）；其次，目标序列与该蛋白质的同源性足够高，一般 50%以上的同源性能够得到比较合理的构象。同源建模一个重要的理论依据就是蛋白质三级结构比氨基酸序列更保守。在生物体中即使差别比较大的同源蛋白也可以被折叠成类似的三维结构，发挥相似的功能。这样的规律也有例外，例如，突变小于 50%序列的一个蛋白质就会折叠成不同的结构。

同源建模的过程一般包括四步：模板选择、序列比对、构建模型和模型评估。前两个过程经常是一起进行的，模板选择通常的方法要基于序列比对的结果。然而，由于数据库搜索技术优先考虑速度而不是准确度，使得比对结果并不一定完全正确，这些过程可以重复进行来改善最终模型的质量。在大规模的自动结构预测中，优化过程的速度和准确度对于结构基因组学的研究很重要，因为数据量太大不可能手工操作，而且结构基因组学的目标是为那些不是结构预测专家的研究人员提供合理正确的模型。

#### 4.3.2.1 模板选择和序列比对

同源建模的第一步是找到最好的模板结构，最简单的方法是使用基于一系列双序列比对的数据库搜索技术，如 FASTA、BLAST，以及敏感度更高的基于多重序列比对的方法，如 PSI-BLAST，通过重复更新特定位点的打分矩阵发现更多的远源同源模板。这类方法能查找出很多的潜在模板，并确定序列的更好模板。蛋白质线索技术（protein threading）在传统同源建模方法中也用于发现模板的搜索技术。

通过这些方法通常会得到很多候选的模板结构，虽然一些方法能从许多模板中获得多杂合模型，但是大部分的方法还是基于单一模板。因此，模板的选择对于最终结构的准确性就很重要。选择模板时可以考虑几个因素，比如目标和模板序列的相似度、功能的差别、模板和目标蛋白的二级结构等。这样，序列比对的叠合区和得到的模型就很重要，有时一个序列使用几个同源的模型，在最后一步才会选择最好的模型。

#### 4.3.2.2 建立模型

用来建立模型的方法主要有三种——序列拼接、片段匹配和空间限制。序列拼接的方法基于完整模型从同源性的保守结构拼接而成，例如，一个哺乳动物中丝氨酸蛋白酶的建模发现核心区的结构相当保守，但是有大部分不同序列的环形可变区则有很大差异。因此，建模时可以先建立保守的核心区，再从已知结构的蛋白中替换可变区的结构。这种方法的关键在于怎么处理非保守区和缺少模板的区域，而可变区的建立通常依靠已知结构的片段文库。

片段匹配的方法是把目标蛋白分成一系列的段片段，每一个片段都与

PDB 中的对应模板相匹配。由于蛋白质结构中的大多数六肽片段都属于仅 100 个左右的不同的结构家族。所以，利用一个被选定的模板结构中的原子位置的子集作为引导位置，通过不断地寻找和组合符合引导结构的片段，可以对目标蛋白做同源建模。因此，片段匹配法中序列比对是分片段进行的，而不是整个蛋白质。每个片段模板的选择基于相应序列的相似性、α 碳坐标的比较、目标蛋白和模板蛋白间不同原子范德华力所造成的空间位阻等。

目前空间限制法也是常用的方法之一。该方法首先从 NMR 得到的实验数据中提取构建蛋白质三维结构所需的数据；然后，用一个或多个目标模版建立一套几何结构标准；接着把这套标准转化成概率密度函数，用于限制新建蛋白质的蛋白质骨架距离和二面角。同时，这套标准也用于对蛋白质中所有重原子进行全局结构最优化。空间限制法对螺旋和折叠的预测表现出色，因此得到了广泛应用，但是这种方法仍然没有很好地解决蛋白质环区的结构预测，这是由于蛋白质环区在溶液中非常柔软、构象多变，目标序列和模板蛋白之间很难有一个精确的匹配。目前流行的同源建模软件 MOLELLER（http：//salilab. org/modeller/）和同源建模数据库 ModBase（http：//modbase. compbio. ucsf. edu/modbase-cgi/index. cgi）使用的就是空间限制法。

#### 4.3.2.3 环区域的建模

目标序列中与模板没有叠合区的区域都是通过环区域建模构建的，在目标蛋白和模板蛋白有很小的序列相似性时最容易发生建模错误。这些未匹配的区域比那些有模板结构对应的区域大体上缺少准确率，尤其是大于 10 个碱基的环。最先的两个侧链二面角（X1 和 X2）对于一个准确的骨架结构可以估计为 30°；然而，接着在长侧链（如赖氨酸和精氨酸）的二面角则很难预测。而且，X1 的小错误就有可能导致侧链末端原子位置的很大偏移，这些原子一般功能都很重要，特别是位于活性中心附近。

#### 4.3.2.4 模型评估

没有真实目标蛋白结构参考时，同源建模结果的评估一般有两种方法：统计势能和基于物理能量计算。两种方法都是对模型的能量做出评估，但是都不能完全符合真实的结构，尤其是 PDB 中数量很少的蛋白，如膜蛋白。

统计势能是一种经验方法，通过 PDB 中已知结构蛋白残基间的相互碰撞频率计算。给两氨基酸相互作用时分配一种可能性或能量值，然后合并成完整模型的能量值。其中有些方法对两残基间的评估效果很差，但是整体模型有合理的得分，而且认为疏水核心和极性氨基酸一般形成球状。比较常用的方法有 Prosa、DOPE 等，该方法比能量计算效率更高。

基于物理能量计算的方法需要获得溶液中稳定蛋白的内部原子相互作用力，特别是范德华力和静电作用。这些计算通常使用分子力场表示，这是因为这些体系对于半经验量子力学方法来说仍然太大了。利用物理能量计算来评估蛋白的方法基于蛋白折叠的能量面假说，该假说认为处在自然状态（native state）下的蛋白质对应于能量面的极小值。在计算能量过程中通常采用隐式溶剂化模型来估算蛋白质分子在溶液中的状态。模型评估中使用的力场称为有效力场（effective force field，EFF），是一种根据 CHARMM 力场原子参数为模型评估专门建立的力场。

模型评估可以使用专门的评估软件（如“What Check”）完成。该软件是荷兰内梅亨大学免费提供的，它可以对模型的接近 200 个方面进行评估。此外，这一软件也可以用来对实验获得的生物大分子结构进行验证。它的网址为：http：//swift. cmbi. ru. nl/gv/whatcheck。

当蛋白质结构已知时，同源建模准确性的评估就很简单，最普通的方法就是使用根均方误差（root mean square deviation，RMSD）度量，在两个结构叠合后计算相应原子的平均距离。然而，RMSD 对于核心区准确但是可变环区不准确的模型评估不足。在 CASP 模型评估实验中引入的新方法称为全局距离测量（global distance test），测出所有原子在某截距时模型和实验结构的距离。这些方法可以用于结构中的所有原子，但通常用于 $\alpha$ 碳和蛋白质骨架原子，这样能够降低侧链旋转所造成的影响。

同源建模还可以用于鉴别相关蛋白间的细微差别，如可以鉴别 $Na^+/K^+$-ATPase 的阳离子结合位点，也可以预测结合强弱。与分子动力学模拟一起，同源建模还能预测一个蛋白质的动力学特征，如钾离子通道的离子选择性。在基因组的蛋白编码区的大规模自动建模已开始在酿酒酵母菌中进行研究，结果产生近 1000 个未知结构的蛋白质模型，也发现 236 个酵母蛋白间新的相互关系。目前常用的同源建模软件有 SWISS-MODEL（http：//swissmodel. expasy. org）、ROBETTA（http：//robetta. org）、FoldX（http：//foldx. crg. es）等。

### 4.3.3 同源建模实例

细胞色素酶 CYP450 是一个大的蛋白质家族，其成员主要负责生物体内药物及小分子的代谢，在多种生物体内均有分布。上海交通大学生命科学与技术学院生物信息学实验室魏冬青教授课题组对该家族蛋白进行了全面的研究，发现该家族中多数蛋白质序列已经测定且同源性很高，但是只有少数几个存在晶体结构。因此，可以应用同源建模方法建立出非常可靠的模型。2004 年首次报道了 CYP2C9 的晶体结构。魏教授课题组顾若虚等以此作为模版对 CYP2C19 进行了同源建模。具体方法如下所述：

CYP2C9 的蛋白序列含有 492 个氨基酸，而 CYP2C19 的蛋白序列含有 490 个氨基酸。顾若虚等通过课题组内部有自主知识产权的软件 SAMM (Shanghai Molecular Modeling) 对 CYP2C9 和 CYP2C19 的蛋白质序列进行了序列比对。他们发现，序列比对的一致性得分与比对的方式密切相关。当以蛋白质的二级结构作为约束条件的时候，比对得分为 52%。如果去掉这个约束条件，纯粹基于序列做比对的时候，得分增加到 91%，这个结果和 NCBI BLAST（http：//www. ncbi. nlm. nih. gov/blast/）的结果相似。考虑到两个酶的结构和功能相似，因此二级结构对酶最终的结果影响很大，他们采取了以二级结构作为约束条件时得到的结果。序列比对结果如图 4.3 所示。

根据 CYP2C9（PDB code 1R90）的晶体结构和序列比对的结果，他们通过片段匹配的方法构建了 CYP2C19 的三维结构模型。之所以使用片段匹配的方式构建，是因为这种方法已经非常成熟，而且有不少成功的例子。例如，在 caspase-3 的晶体结构被测定之前，曾经用片段匹配方法构建 caspase-8 蛋白酶结构域的三维结构模型。为了得到一个合适的模板，首先以 caspase-1 的晶体结构为模板建立了 caspase-3 催化结构域的三维结构模型，然后以得到的三维结构为模板，进一步构建了 caspase-8 蛋白酶结构域的三维结构模型。之后的实验测定的晶体结构显示，计算机构建的 caspase-3 和 caspase-8 的三维结构模型与它们的晶体结构非常相似，比如：caspase-3 蛋白酶结构域的晶体结构和其计算机结构模型的所有骨架原子位置的标准偏差（RMSD）为 2.7Å，而 caspase-8 蛋白的相应值为 3.1Å，其核心结构的标准偏差仅为 1.2Å。随后，许多蛋白质都用片段匹配的方法成功建立了三维结构模型，如：以 RAIDD CARD 的核磁共振结构为模板，建立了 Apaf-1、Ced-4 和 Ced-3 的 CARDs（caspase recruitment domains）结构域的三维结构模型；此外还有 Cdk5-Nck5a* 复合体和 caspase-9 蛋白酶结构域的三维结构模型。

Cdk5-Nck5a* 复合体的三维结构的计算机模型发表之后两年，它的晶体结构被测定了，人们发现二者基本上是相同的。Cdk5 和 Nck5a* 相结合时被包裹的表面积为 3400Å$^2$，而通过计算机模拟得到的结果为 3461Å$^2$。同时，基于 Cdk5-Nck5a* -ATP 的计算机结构模型，人们进行了分子截短的实验，结果确认并拓展了计算机模型所做出的预言。片段匹配方法还被用于构建 beta-分泌酶酶原的三维结构模型并进一步为 beta-分泌酶的前体不能像 Shi 和 Benjannet 等所提到的普通酶原那样抑制活性提供了令人信服的解释。

最近片段匹配的方法被用于构建许多蛋白质的细胞外结构域的三维结构模型，如：1，2，3，5 四种亚型的 GABA-A 受体，BACE2，H5N1-NA，

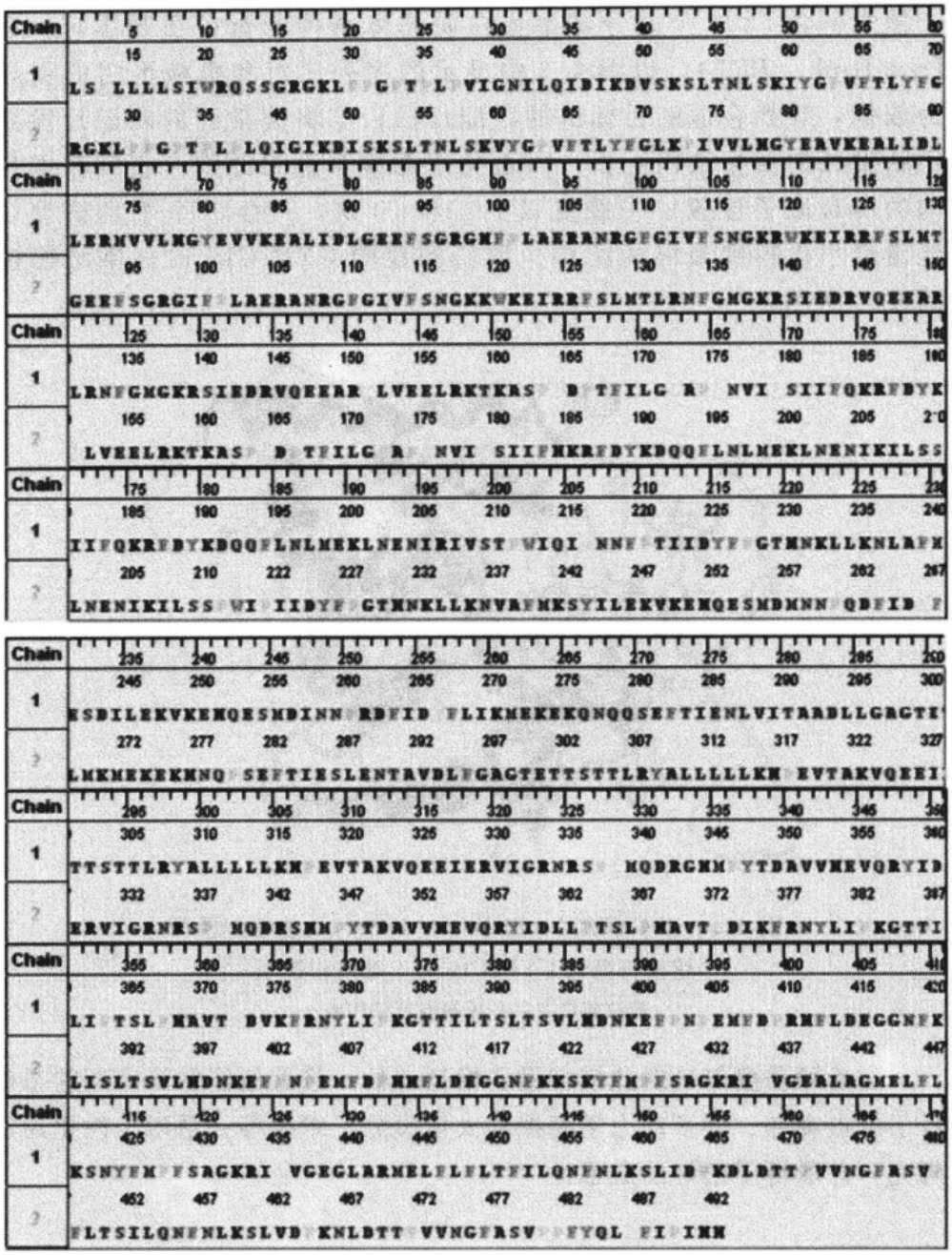


图 4.3　CYP2C19（Chain 1）和 CYP2C9（Chain 2）的序列比对（见彩图）
其中氨基酸按其不同的功能标记为不同的颜色：红色，蓝色，粉色，深绿，浅绿，黄色，橙色。分别代表该氨基酸为酸性的，碱性的，中性亲水的，脂肪族，芳香族，含硫的，含亚氨基的两条序列的相似度为 52%

GFAT，cathepsin-E，alpha-7 烟碱乙酰胆碱受体和人类的钾通道和钠通道蛋白。

顾若虚等的研究以 CYP2C9（PDB code 1R90）为模板经过以下四步

完成了对 CYP2C19 酶的建模：①把目标链断裂成一系列短的片段；②搜索包含 5200 个高分辨率蛋白质晶体结构的蛋白质数据库（Protein Data Bank，PDB），以图 4.3 中显示的序列比对和模板蛋白的引导位置为依据，寻找合适的目标链的匹配片段；③将搜索到的匹配片段添加到不断增长的目标结构上去，以避免范德华重叠为原则，直到获得目标结构的所有原子位置；④重复以上三步 10 次，产生一个平均模型，通过能量最小化得到最终的结构。他们构建的 CYP2C19 的三维结构模型如图 4.4 所示。

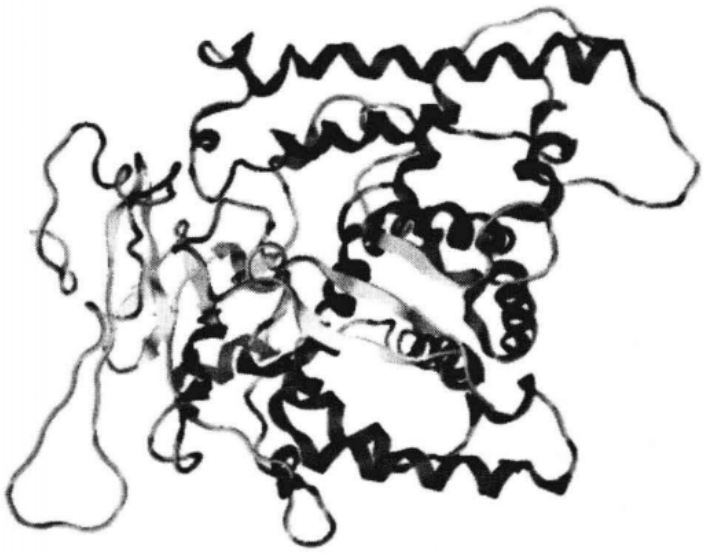


图 4.4　以 CYP2C9（PDB code 1R90）为模板作同源建模得到的 CYP2C19 的三维结构模型
蛋白质骨架以平板和缎带结构表示

随后顾若虚等又对此模型做了模型评估。接着他们在该模型的基础上进行了单点突变、分子动力学模拟和大规模分子对接等一系列个性化药物设计研究，并取得了很好的成果。

## 4.4　蛋白质四级结构预测

蛋白质的四级结构是由不同结构域相互组合而成的。四级结构的预测就是对蛋白质-蛋白质相互作用方式的预测。它是生物信息学和结构生物学的综合应用。这种预测主要分为两大类。一类从系统发生和蛋白质共进化的角度出发，应用数学模型、统计学模型来预测蛋白质的相互作用。这种方法可以预测的蛋白质很多，不论蛋白质有没有结构都可以预测，但是不能给出两个蛋白质的结构信息和结合方式。另一类是从两个蛋白质结构出发，利用物

理学原理进行的预测，又称为蛋白质对接（protein-protein docking）或者大分子对接（macromolecular docking）。这种方法通过计算两个蛋白质不同结合模式下的形状互补性、结合能/自由能等特征，来判断该结合模式的优劣。结合能的计算主要根据力场参数（force field）进行，优劣的综合判断依靠打分函数（scoring function）实现。

四级结构的预测还不成熟，需要进一步的发展。目前可以进行四级结构预测的工具有 STRING（http：//string-db.org）、PPIs（http：//www.compbio.dundee.ac.uk/www-pips）、metaPPI（http：//scoppi.biotec.tu-dresden.de/metappi）、ZDOCK（http：//zdock.bu.edu）等。

## 4.5 结束语

蛋白质结构预测意义非常重大。从 20 世纪 60 年代开始产生到现在更发展成为一个生物信息学、计算机科学、数学、物理和化学等多学科交叉的热点研究领域。原有的计算方法不断改进，全新的方法、思路不断涌现，随之而来的是预测精度和深度的不断提高。然而，纵观学科发展，现有的预测方法仍然不能满足社会应用，特别是制药领域和酶工业对蛋白质结构信息的需求。正因为如此，经过半个世纪的发展，这个学科仍然具有非常广阔的空间等待人们去探索。

## 参考文献


[1] Dor O and Zhou Y. Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training. Proteins: Structure, Function, and Bioinformatics, 2007, 66 (4): 838-845.

[2] Chou P Y and Fasman G D. Prediction of protein conformation. Biochemistry, 1974, 13 (2): 222-245.

[3] Mount D W. Bioinformatics: sequence and genome analysis. CSHL press, 2004.

[4] Garnier J, Osguthorpe D J, and Robson B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. Journal of Molecular Biology, 1978, 120 (1): 97-120.

[5] Kloczkowski A, et al. Combining the GOR V algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence. Proteins: Structure, Function, and Bioinformatics, 2002, 49 (2): 154-166.

[6] Sen T Z, et al. GOR V server for protein secondary structure prediction. Bioinformatics, 2005, 21 (11): 2787.

[7] Cole C, Barber J D, and Barton G J. The Jpred 3 secondary structure prediction server. Nucleic acids research, 2008, 36 (suppl 2): W197-W201.

[8] Pham T H, Satou K, and Ho T B. Support vector machines for prediction and analysis of beta and gamma-turns in proteins. Journal of bioinformatics and computational biology, 2005, 3

(2): 343-358.

[9] Zimmermann O and Hansmann U H E. Support vector machines for prediction of dihedral angle regions. Bioinformatics, 2006, 22 (24): 3009.

[10] Karypis G, YASSPP: better kernels and coding schemes lead to improvements in protein secondary structure prediction. Proteins: Structure, Function, and Bioinformatics, 2006, 64 (3): 575-586.

[11] Samudrala R and Moult J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. Journal of Molecular Biology, 1998, 275 (5): 895-916.

[12] Bonneau R and Baker D. ABI NITIO P ROTEIN S TRUCTURE P REDICTION: Progress and Prospects. Annual Review of Biophysics and Biomolecular Structure, 2001, 30 (1): 173-189.

[13] Sanchez R. Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct, 2000, 29: 291-325.

[14] Dalal S, Balasubramanian S, and Regan L. Transmuting [alpha] helices and [beta] sheets. Folding and Design, 1997, 2 (5): R71-R79.

[15] Dalal S, Balasubramanian S, and Regan L. Protein alchemy: Changing β-sheet into α-helix. Nature Structural & Molecular Biology, 1997, 4 (7): 548-552.

[16] Baker D and Sali A. Protein structure prediction and structural genomics. Science's STKE, 2001, 294 (5540): 93.

[17] Wallner B and Elofsson A. All are not equal: a benchmark of different homology modeling programs. Protein Science, 2005, 14 (5): 1315-1327.

[18] Levitt M, Accurate modeling of protein conformation by automatic segment matching* 1. Journal of molecular biology, 1992, 226 (2): 507-533.

[19] Fiser A and Sali A. ModLoop: automated modeling of loops in protein structures. Bioinformatics, 2003, 19 (18): 2500.

[20] Sippl M. Recognition of errors in three-dimensional structures of proteins. Proteins: Structure, Function, and Bioinformatics, 1993, 17 (4): 355-362.

[21] Lazaridis T and Karplus M. Discrimination of the native from misfolded protein models with an energy function including implicit solvation 1. Journal of molecular biology, 1999, 288 (3): 477-487.

[22] Ogawa H and Toyoshima C. Homology modeling of the cation binding sites of $Na^+/K^+$-ATPase. Proceedings of the National Academy of Sciences, 2002, 99 (25): 15977.

[23] Capener C, et al. Homology modeling and molecular dynamics simulation studies of an inward rectifier potassium channel. Biophysical Journal, 2000, 78 (6): 2929-2942.

[24] Sánchez R and Sali A. Large-scale protein structure modeling of the Saccharomyces cerevisiae genome. Proceedings of the National Academy of Sciences of the United States of America, 1998, 95 (23): 13597.

[25] Pellegrini M, et al. Assigning protein functions by comparative genome analysis protein phylogenetic profiles. 2003, Patents. US 6564151B1.

[26] Pazos F and Valencia A. Similarity of phylogenetic trees as indicator of protein-protein interaction. Protein Engineering Design and Selection, 2001, 14 (9): 609.

# 第5章 蛋白质结构模拟在蛋白研究中的应用实例

王 莹 王靖方 魏冬青

## 5.1 引言

蛋白质结构模拟主要是基于分子模拟理论。简单的分子模拟可以采用机械模拟或者是用笔和纸做计算。然而，生物体中的蛋白质结构相对于简单的分子是相当复杂的，从前的方法难以将蛋白质体系的运动状态和性质描述清楚。电子计算机的出现及其相关技术的发展，解决了这一难题，使蛋白质模拟进入了一个前所未有的光明时代。因此，现在科学家们将量子力学、分子力学、构象分子以及其他的利用计算机来理解和预测分子体系的方法结合在一起，来观察、描述和模拟蛋白质分子体系的行为。

在蛋白质模拟的过程当中，经常会提到“模型”一词，所谓的模型就是用最简单的方式使人们形象地认知分子的三维结构。最为著名的模型如Dreiding等发展的球棍模型，Corey、Pauling和Koltun等发展的CPK模型等。这些模型广泛地应用于教学和科研工作之中。蛋白质结构模拟研究可以概括为以下三个过程。第一步是建立一个模型用于描述蛋白体系的分子间和分子内的相互作用。在模拟研究中最常用的两种模型是量子力学和分子力学模型，利用它们可以计算出体系中原子和分子的任意一种分布的能量，并且可以观察到随着体系中原子或分子位置的改变，体系能量是怎样变化的。第二步是进行能量最小化、分子动力学模拟、蒙特卡罗模拟和构象搜索等一系列体系自身内的计算。最后一步就是将以上两步得到的计算结果进行分析。

本章主要是通过列举实例，让大家进一步理解这三个步骤的具体操作过程，为以后灵活运用蛋白质结构模拟这种方法来研究大家感兴趣的蛋白质提供帮助。

## 5.2 蛋白质结构的预测

研究蛋白质的空间结构意义重大，研究蛋白质结构，有助于了解蛋白质的作用，了解蛋白质如何行使其生物学功能，认识蛋白质之间或蛋白质与其他分子之间的相互作用都非常重要。而且，在所有的蛋白质模拟研究中，蛋白质结构是基础，所有的计算都起始于蛋白质的结构。所以，如何得到准确的蛋白质结构，是正确模拟蛋白质的必要条件。

目前蛋白质结构的测定技术主要有 X 射线晶体衍射法（X-ray diffraction）、二维和多维核磁共振技术（nuclear magnetic resonance，NMR）。通过这些技术测得的生物大分子三维结构，都收集在蛋白质结构数据库（Protein Data Bank，PDB）中，通过这个数据库我们可以轻松地获得蛋白质的 PDB 结构文件并进行计算。尽管蛋白质结构测定技术有了较为显著的进展，但是通过实验方法确定蛋白质结构的过程仍然非常复杂，代价较高，因此实验测定的蛋白质结构比已知的蛋白质序列要少得多。所以出现了许多蛋白质结构的预测方法，主要包括同源建模法、折叠识别法、从头计算法等。这些方法的原理之前已经有所介绍，我们就不在这里赘述了。由于同源建模法是目前应用最为广泛的蛋白质三维结构预测方法，所以我们在这里就以同源性方法为实例，讲解如何预测蛋白质的三维结构。

一般同源建模法推测蛋白结构需要以下六个步骤：第一步模板蛋白的搜索；第二步目标蛋白质和模板蛋白的序列比对；第三步主链结构建模；第四步环区建模；第五步侧链建模和优化；第六步整体结构优化。这些步骤可以用不同的程序分别完成，也可以使用同源建模服务器或者自动建模的程序来完成。现在，有很多在线工具也可以实现这一过程，比如 SWISS-MODEL、MODELLER、PARMODEL、HHpre、MODBASE、3D-JIGSAW、PROTEUS 和 I-TASSER 等。得到的推算模型还需要进行评估，以确定其优劣。Very3D 和 Procheck 是两种最为常用的用来评估结构的打分函数。

随着同源建模方法的日趋成熟，运用一些自动建模的程序也可以得到很准确的蛋白质三维结构。SWISS-MODEL 是一种可以通过 ExPASy 服务器的 Web 界面（http：//swissmodel. expasy. org）免费访问的自动化蛋白质同源建模服务器。首先，SWISS-MODEL 的氨基酸序列提交方式分为三种：

（1）简洁模式（first approach mode） 可以在 Web 界面直接贴入预测蛋白质的氨基酸序列或是 Swiss-Prot/TrEMBL 编号，然后服务器可以完全自动地建立模型。用户也可以指定模板结构。

（2）联配模式（alignment mode） 这个模式需要多序列联配的方式，

用户需要提供一条目标序列和一条模板序列。

(3) 项目模式 (project mode) 这种模式适合于对建模有细节要求的用户，此模式可以选择不同的模板，手工编辑目标序列和模板的联配结果，以便判断插入和删除的位置。项目模式还可以用于重复改进简洁模式的结果，而且可以用于构建复合蛋白的模型。

**SWISS-MODEL 的结构与评估** SWISS-MODEL 将通过邮件输出模型的坐标，可以是 DeepView (是一种整合工具，用于观察和分析蛋白的结构) 项目文件的方式或者是普通的 PDB 文件格式。在结果的日志文件中还包含了一个目标模型的变异参数 C-score，当模型中的某些部分找不到匹配的模板信息时，此值显示为 99，而且日志文件还记录了整个结构以及每个残基的力场能量，还可以选择在结构中包含一份蛋白校验工具 WhatCheck (http：//www. cmbi. kun. nl/gv/whatcheck) 的报告。另外，也可以利用服务器上的 ANOLEA 原子平均势能的 Web 界面 (http：//swissmodel. expasy. org/anolea) 检查模型的质量。

## 5.3 分子对接

分子对接是依据配体与受体“锁-钥原理”，模拟小分子配体与受体生物大分子之间的相互作用。配体与受体相互作用主要包括静电作用、氢键作用、疏水作用和范德华相互作用等。通过对空间匹配和能量匹配的计算，预测两者之间的结合模式和亲和力，从而得到最优构象。其中空间匹配的计算，通常采用格点计算和片段生长等方法，能量计算则使用人工神经网络、禁忌搜索、局部搜索、模拟退火或遗传算法等方法。各种分子对接方法对体系均有一定的简化，根据简化的程度和方式，将分子对接方法分为三类：刚性对接，半柔性对接和柔性对接。可将这些分子对接方法运用于小分子药物的虚拟筛选，或分子动力学模拟前的结构准备。

目前，文献报道的分子对接软件很多，这些软件中很多都是由实验室开发，免费试用，当软件逐渐完善没有缺陷时，就被商业公司购买，作为某大型软件包中的模块。当然，也有很多软件依然免费提供强大的服务，甚至是源代码 (source code)。所以，接下来我们主要介绍几款免费又好用的分子对接软件。

(1) AutoDock 最新数据显示，AutoDock 是引用文献最多的软件，由 Olson 科研组开发并向外提供源代码的免费软件 (http：//autodock. scripps. edu/)。AutoDock 应用的是半柔性对接方法，允许小分子的构象发生变化，以结合自由能作为评价对接结果的标准。从 AutoDock3. 0 开始采用拉马克遗传算法 (LGA)，这种方法将遗传算法与局部搜索方法相结合，

先用遗传算法搜索势能面，再利用局部搜索精细优化势能面。最近AutoDock又开发出一种可以在Windows环境下运行的AutoDock-Vina版本，其特点是速度快，与AutoDock4.0相比更具有优势。

(2) DOCK　说起分子对接就不能不说DOCK这款软件，它也是以源代码发布的免费软件（http://docking.org/），是由Kuntz课题组开发。DOCK应用半柔性对接法，固定小分的键长和键角，将小分子拆分成若干刚性片段，根据受体的表面几何性质，将小分子的刚性片段重新组合，进行构想搜索。同时，考虑分子之间的静电相互作用、范德华力等非键相互作用，搜索势能面，最终用能量评分和原子接触罚分之和对对接结果进行评价。

(3) Surflex-Dock　Surflex-Dock是对学术用户免费的商业软件（http://www.biopharmics.com/products.html），由加州大学旧金山分校Ajay N. Jain教授开发。这种方法的评分函数来自于已知的结合数据，使用阴性训练数据以减少假阳性结合分数，还可以进行小分子复合物的叠合。

(4) FlexX　FlexX是另一种运算速度超快的分子对接软件，是由德国国家信息技术研究中心生物信息学算法和科学计算研究室开发的分子对接软件（www.biosolveit.de/FlexX/），目前已经作为分子设计软件包Sybyl的一个模块实现商业化。FlexX使用碎片生长的方法寻找最佳构象，根据对接自由能的数值选择最佳构象。

还有许多其他的对接软件，主要对接软件列于表5.1，在此就不一一介绍了。以上讲解的是进行大分子与小分子对接的软件，还有一些软件可以做大分子之间的分子对接工作，比如ESCHER、MONTY、HADDOCK、Z-DOCK和HEX等软件。这些软件还在进一步改善过程中，所以我们不作过多介绍。

**表5.1　大分子与小分子对接软件列表**

| 软件名称 | 收费形式 | 运算速度 | 主　页 |
|---|---|---|---|
| AutoDock | 免费 | 一般 | http://autodock.scripps.edu/ |
| DOCK | 免费 | 快 | http://docking.org/ |
| 3D-DOCK | 免费 | 慢 | http://www.bmm.icnet.uk/docking/ |
| FRED | 学术用户免费1年 | 快 | http://www.eyesopen.com/forms/eval_request.php |
| Surflex-Dock | 学术用户免费 | 快 | http://www.biopharmics.com/products.html |
| FlexX | 收费，可申请6周试用版 | 快 | http://www.biosolveit.de/FlexX/ |
| Glide | 收费 | 一般 | http://www.schrodinger.com/products/14/5/ |
| GOLD | 收费 | 快 | http://www.ccdc.cam.ac.uk/products/life_sciences/gold/ |
| ICM-Dock | 收费 | 快 | http://www.molsoft.com/index.html |
| MVD | 1个月的试用期 | 慢 | http://molegro.com/mvd-trial.php |

接下来我们还是通过实例给大家以直观的印象，进一步了解如何使用这

些软件。AutoDock 作为一款免费且功能强大的软件深受科研人员的喜爱，所以我们以 AutoDock 为例，介绍分子对接的整个过程。

AutoDock 对接过程主要包括以下几个步骤：准备生物大分子受体→创建配体小分子文件→准备栅格图→设置对接参数文件→运行 AutoGrid 和 AutoDock→观察对接结果。

a. 准备目标蛋白　用 pyMol 等蛋白分析软件打开蛋白的 PDB 文件，这里以 hsg1. pdb 为例，移走所有的水分子后保存。在 AutoDockTool（ADT）中，打开 hsg1. pdb 文件。

File→Read Molecule 选择 hsg1. pdb 文件。

Edit→Hydrogens→Add 选择“Polar only”“noBondOrder ”“yes”，加入极性氢离子。

File→Save→Write PDB 保存文件。

b. 创建配体小分子文件

Ligand→Open…（AD3）打开小分子文件。

Display→Show/Hide Molecule 将 macromolecule 隐藏，放大 Ligand。

Ligand→Torsion Tree→Detect Root…程式判定的 root 会以一颗绿球表示。

Ligand→Torsion Tree→Choose Torsions…打开“Torsion Count”对话框，选择哪些键可以转动，哪些键不可以转动。绿色代表可以转动的键，红色代表固定键，粉色代表设定的不可转动的键。

Ligand→Output→Save as PDBQ…（AD3）保存小分子 PDBQ 文件命名为“Ligand. pdbq”。

c. 准备栅格图

Grid→Macromolecule→Choose…（AD3）点击“Select Molecule”，选择第一步准备的目标蛋白质文件并储存为“Protein. pdbq”。

Grid→Set Map Types→Choose Ligand…（AD3）Select the ligand，会跳出一个“AutoFpf Ligand”对话框，点击“Accept”。

Grid Box…打开“Grid Options”对话框，可以调整 Grid Box 的大小、grid points 之间的距离（Spacing）和中心位置（Center Grid Box），调整完毕后关闭对话框：File→Close Saving Current。

Grid→Output→Save GPF…（AD3）保存为“Grid. gpf”文件。

d. 设置对接参数文件

Docking→Macromolecule→Set filename…（AD3）选择 Protein. pdbq 文件。

Docking→Ligand→Choose…（AD3）选择小分子文件“Ligand. pdbq”，同时跳出一个对话框，无需修改，保持默认值，点击“Accept”。

Docking→Search Parameters→Genetic Algorithm… 选择遗传算法。

Docking→Search Parameters 选择包括 random number generator，random number generator seeds，the energy outside the grid 等。如 Number of runs→50；Population size→50；Maximum number of energy evaluations→1500000；Translation→0.2；Quaternion→5.0；Torsion→5.0；RMS Cluster Tolerance→1.5，其他值保持默认值。

Docking→Output→Lamarckian GA…（AD3）保存这个对接文件为“Dock.dpf”。

e. 运行 AutoGrid 和 AutoDock

autogrid3→Grid.gpf→Grid.glg 并运行 AutoGrid3。

autodock3→Dock.dpf→Dock.dlg 并运行 AutoDock3。

f. 观察对接结果

可以通过 ADT 观察对接结果：

Analyze→Dockings→Open…打开 Dock.dlg 文件，了解对接自由能的内容。

Analyze→Macromolecule→Choose 加载目标蛋白分子。

Select→Direct Select→Molecule List … 选择大分子。

Color→by Atom Type 设置蛋白质分子原子的颜色。

Compute→Molecular Surface→Compute Molecular Surface 计算大分子表面。

Analyze→Conformations→Play … 观察构象。

Analyze→Clustering→Show … 观察构象簇。

最后根据构象簇的情况和自由能大小来判断最适宜的对接分子或构象，以达到对接的目的。

## 5.4 蛋白结构预测与分子对接在生物研究中的应用

之前的实例分别利用一种分子模拟方法来解决某一方面的问题，但是在很多情况下，例如刚刚发现某一蛋白，对其了解甚少时，就需要综合使用多种模拟方法，深入理解这一蛋白质的结构和功能特性。所以我们以 H5N1 高致病性禽流感病毒蛋白的分子模拟为例，讲解蛋白质结构预测与分子对接在蛋白质研究中的综合运用。

H5N1 是一种高毒力的禽流感病毒，它可以大规模感染禽类，在某些特殊情况下还可以感染人类。据世界卫生组织（WHO）统计，在 2004 年 2 月到 2005 年 11 月期间，禽流感病毒已经传播到 15 个亚欧国家。由于在此项研究报道之前没有任何关于 H5N1 蛋白质晶体结构数据可以利用，但又急

需解决 H5N1 产生的耐药性问题，且了解其蛋白酶的结构与活性之间的关系，所以魏冬青教授等采用了结构生物信息学的方法构建了 H5N1-NA 的三维结构，又模拟了 H5N1-NA 与药物分子对接过程，从而论证了 H5N1 病毒产生耐药性的原因。

（1）构建 H5N1-NA 三维结构　根据同源建模的原理，通过序列比对的方法，运用 PILEUP 程序中的 GCG 软件包，找到一个相似度为 47%的 N9-NA（PDB code：1F8B）蛋白作为模板（序列比对过程见图 5.1），用片段

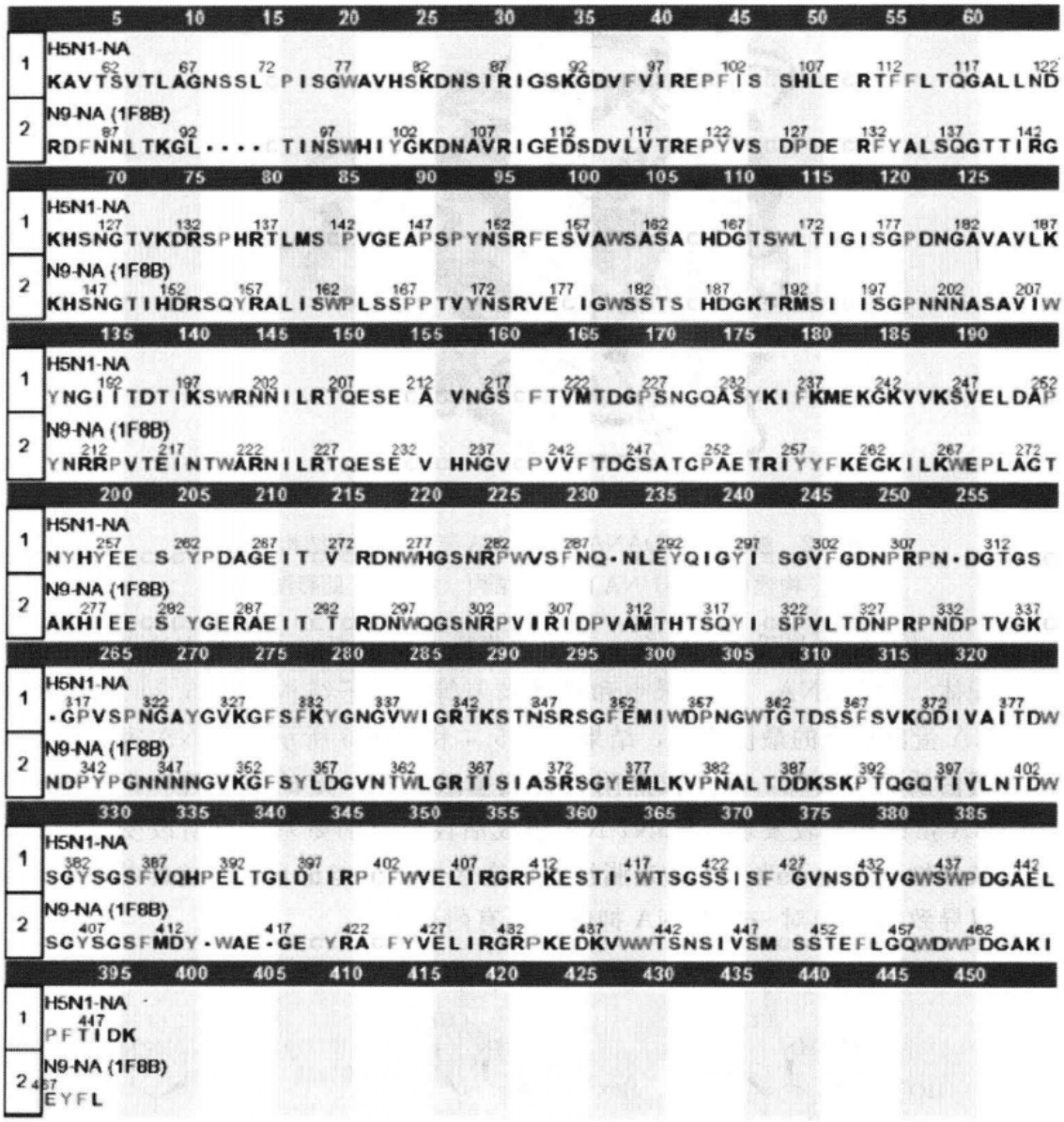



图 5.1　序列 H5N1-NA（Accession No. Q5H89）和 N9-NA（Accession No. Q84070）比对（见彩图）

图中氨基酸颜色表示氨基酸的功能：酸性—红色；碱性—蓝色；中性亲水—粉红色；脂肪族—墨绿色；芳香族—青绿色；包含巯基—黄色；亚氨基—橙色

匹配和坐标重建的方法建立起 H5N1-NA 的三维结构，如图 5.2。从图中可以看出，两个蛋白质的基本骨架叠合得很好，只是由于部分氨基酸片段的插入与删除，使得 Loop 结构有所不同。图中 DANA 所在的位置正是 NA 的活性口袋，这个口袋由 14 个保守的带电荷的氨基酸和 3 个疏水氨基酸组成。进一步分析催化口袋与小分子底物是如何结合的就可以了解 H5N1 的耐药性是如何产生的。

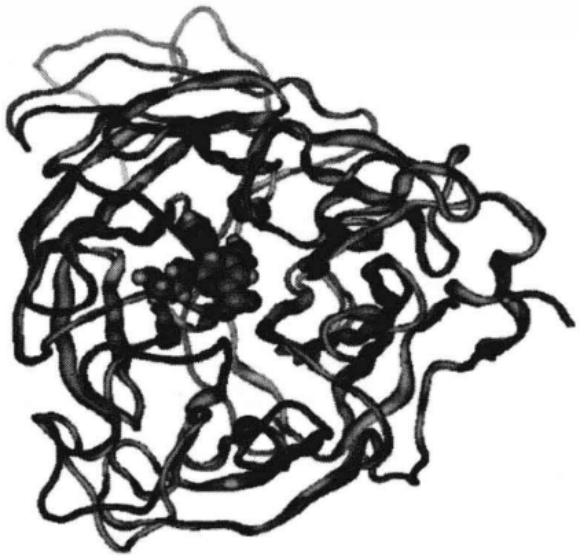


图 5.2　根据结合 DANA 的 N9-NA 复合物晶体结构（红色）构建的 H5N1-NA 的三维结构（绿色）（见彩图）

（2）配体与受体的结合　利用“蒙特卡罗模拟退火”的方法搜索小分子配体，如 DANA、奥塞米韦和扎那米韦等（分子结构见图 5.3）与 H5N1-NA 蛋白结合的最优构象，结果如图 5.4 和图 5.5 所示。与 NA 家族的其他蛋白质比较发现，活性位点附近的氨基酸具有一定的保守性，但是与 N9-NA 蛋白相比较发现，H5N1-NA 组成活性位点的氨基酸有所改变，周围的亲水性也发生了改变，由亲脂性变成亲水性。可能正是由于这样的改变，所以导致 H5N1 对一般的 NA 抑制剂具有耐药性。

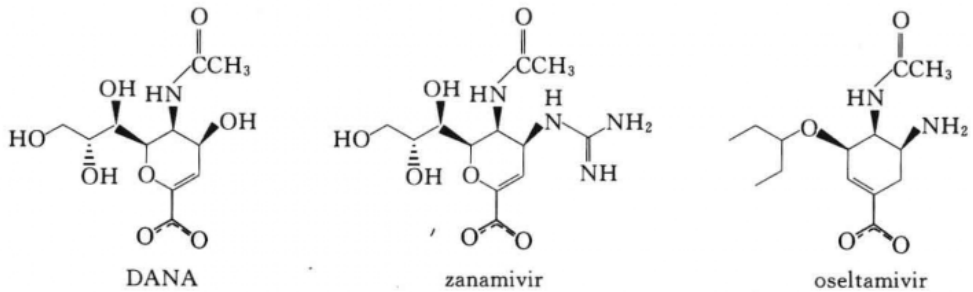
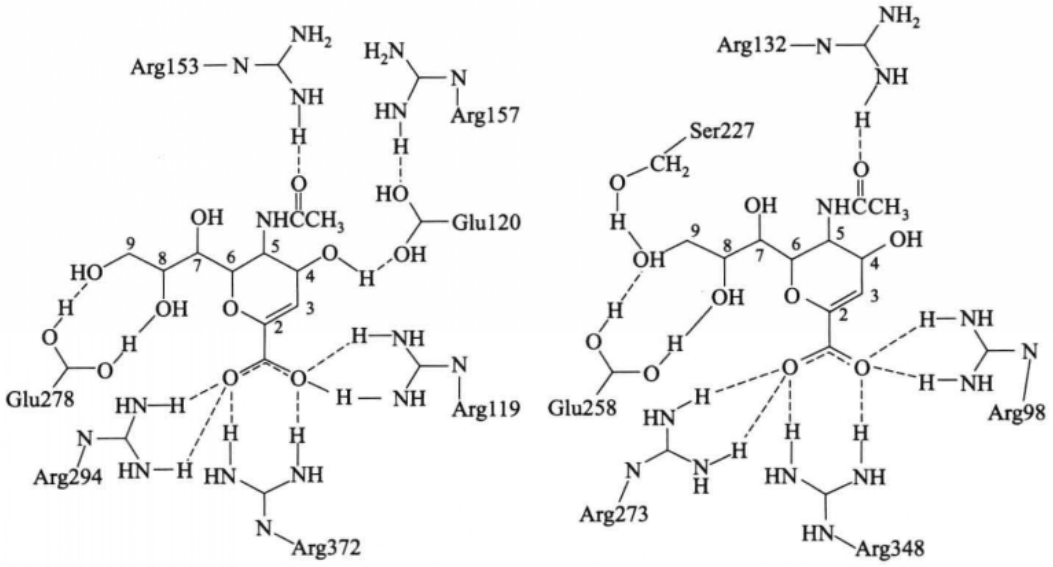
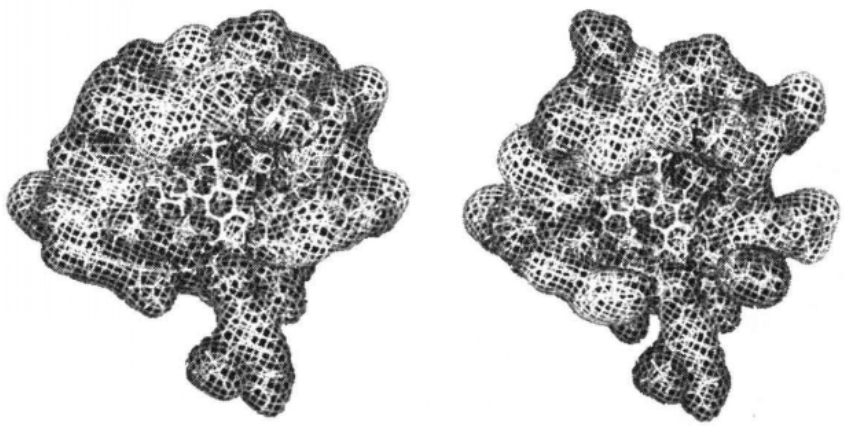



图 5.3　DANA、扎那米韦和奥塞米韦分子结构




(a) N9-NA:DANA　　(b) H5N1-NA:DANA

图 5.4　配体 DANA（红色）分别与受体 N9-NA（a）和 H5N1-NA（b）活性位点残基（蓝色）相互作用示意图（见彩图）



(a)　　(b)

图 5.5　N9-NA（a）和 H5N1-NA（b）活性空腔的亲脂和亲水表面示意图（见彩图）其中绿色表示亲脂性，蓝色表示亲水性

本实例通过蛋白质预测与分子对接方法的联用，实现了对未有晶体结构数据参考的 H5N1-NA 蛋白质结构和功能的研究，并从分子模拟的角度推测 H5N1-NA 的耐药性产生原因，为以后的研究打下可靠的理论基础。

## 5.5　分子动力学模拟在蛋白结构变化中的应用

分子动力学是一门综合物理、数学和化学的交叉学科，它以起始构型为

模拟基础，赋予构成分子的各个原子速度，并按照分子瞬时的运动状态，求解这些原子的牛顿运动方程，得到每个原子新的位置和速度。分子动力学模拟可以作为X射线晶体衍射和核磁共振等实验方法测定蛋白质空间结构的一种补充。这种方法可以用于研究蛋白质等生物大分子体系在生理环境下或者其他特定环境下的空间结构变化。如果在模拟时间内为分子体系选取正确的力场函数形式及其参数，运用分子动力学就可以准确地解析分子运动性质。

分子动力学这门学科已有很久的发展历史，早在1957年Alderand等就提出了基于钢球势的分子动力学法。但是由于计算资源的限制，最近几十年才将分子动力学广泛应用于生物大分子的研究中，并取得了令世人瞩目的成就。分子动力学为分子模拟提供了许多的帮助。运用图形与分子模型交互，可在计算机屏幕上以图形的形式表示出分子动力学的轨迹数据，观察、比较和分析分子各个构象式的三维立体结构，用以推测和评价生物大分子的构象变化或药物与受体相互作用的分子形式。本节以T1脂肪酶的分子动力学模拟研究为实例，向大家展示分子动力学模拟在分析生物大分子结构特点及分子结构变化中的应用。

### 5.5.1 研究背景

T1脂肪酶是从马来西亚的*Geobacillus zalihae*菌株中分离出来的一种分泌于细胞外的恒温噬碱酶。在温度的诱导下，酶的分子结构开始改变，在大约70℃时，此酶具有很高的活性，将长链的甘油三酯催化水解为脂肪酸和甘油。T1脂肪酶含有经典的催化三联体Ser-His-Asp结构，整个蛋白骨架由7条β折叠片被13个α螺旋、10个$3_{10}$单环及Loop包围成球形（见图5.6），实验证明，酶的活性位点埋在分子里面。分子动力学模拟T1脂肪酶的目的就是从微观角度阐述T1脂肪酶在各个温度下结构的变化，以及分子水平的催化机理。

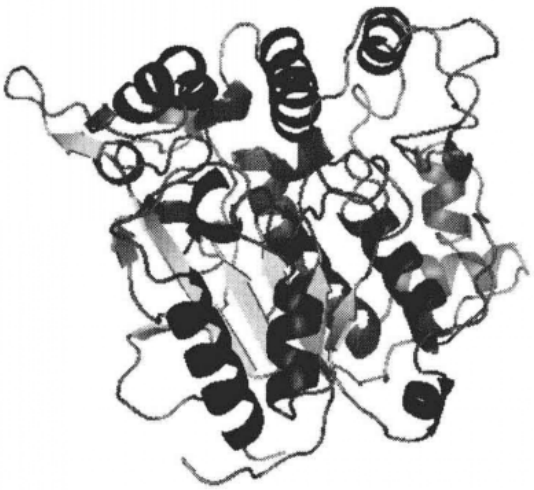


图5.6 T1脂肪酶的分子结构

### 5.5.2 实验方法

以T1脂肪酶的晶体结构（PDB code：2dsn）为初始模型，运用GROMACS 3.3.3软件包进行分子动力学模拟。选用GROMOS 53a6作为整个蛋白的力场，设置周期性边界条件。将蛋白放入8.5nm×8.5nm×8.5nm大小的盒子中，并加入SPC216型的水分子。加入6个钠离子使整个体系达到离

子平衡后，进行 1000 步的能量最小化模拟。最后分别在 30℃、60℃、70℃和 100℃温度条件下，完成 30ns 的分子动力学模拟。在模拟过程中，运用 LINear Constraint Solver（LINCS）算法计算键约束；PME 算法来计算体系中的静电势能；设定范德华势能计算中的 CUTOFF 为 12Å。体系坐标每隔 1ps 记录一次，但分子动力学模拟的时间步长为 2fs。

### 5.5.3 实验结果

通过分析分析动力学模拟 T1 脂肪酶运动的轨迹，我们发现 T1 脂肪酶在低温（30℃）环境下很不稳定，计算轨迹与初始结构间 Cα 的均方根偏差（RMSD）来判断系统是否收敛及结构变化程度，结果见图 5.7，30℃时 RMSD 值较高且收敛得很慢，且 Cα 的回转半径值（见图 5.8）明显高于其他温度。相反，T1 脂肪酶在较高温度下（70℃），Cα 的均方根偏差（RMSD）值收敛得快且低于其他温度，证明此酶在 70℃条件下结构较稳定，有利于催化底物。

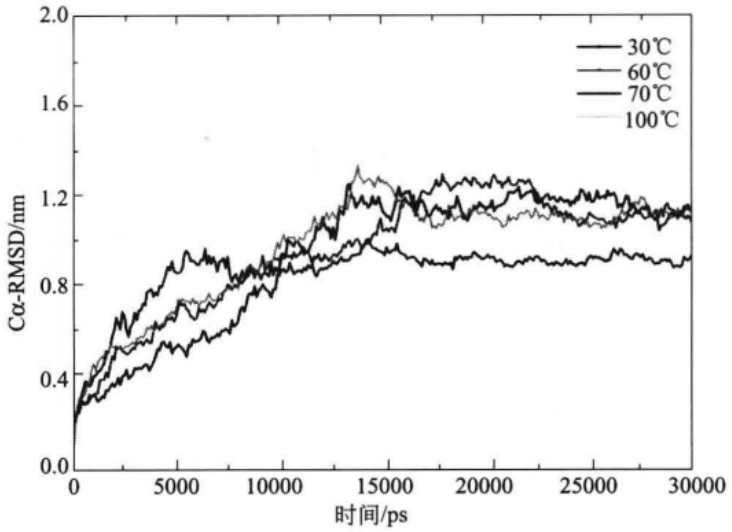



图 5.7 模拟过程中骨架碳原子的均方根偏差随时间变化示意图（见彩图）

用蛋白质结构可视化软件（如 VMD，PyMOL 等）仔细观察 T1 脂肪酶在各个温度条件下的结构变化，我们发现活性位点 Ser112 的上方有两个螺旋结构的运动有一定的规律，所以我们计算这两个螺旋-螺旋 6（Asp175-Ala190）和螺旋 9（Val294-Ser301）在 30ns 的模拟过程中距离的变化（见图 5.9）。由图可知，30℃时，两个螺旋之间的距离变化较大，距离最远可以达到 4.9nm，最近是 2nm。这两个时刻蛋白的结构如图 5.10 所示，分别呈现打开和闭合的结构。而在 70℃时，两个螺旋之间的距离虽也有波动，但是波动幅度明显比 30℃小，且在 10ns 以后，T1 脂肪酶的打开构象更为稳定。如果与底物结合，那么在 70℃时，底物与酶之间的结合时间相对于

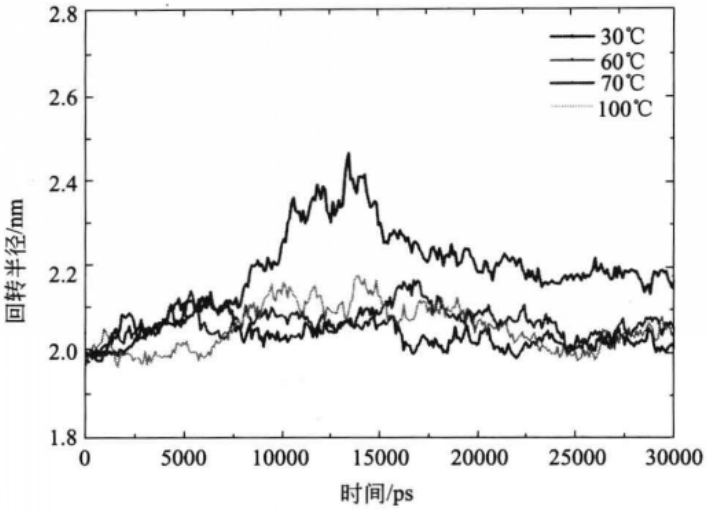



图 5.8　骨架碳原子的回转半径示意图（见彩图）

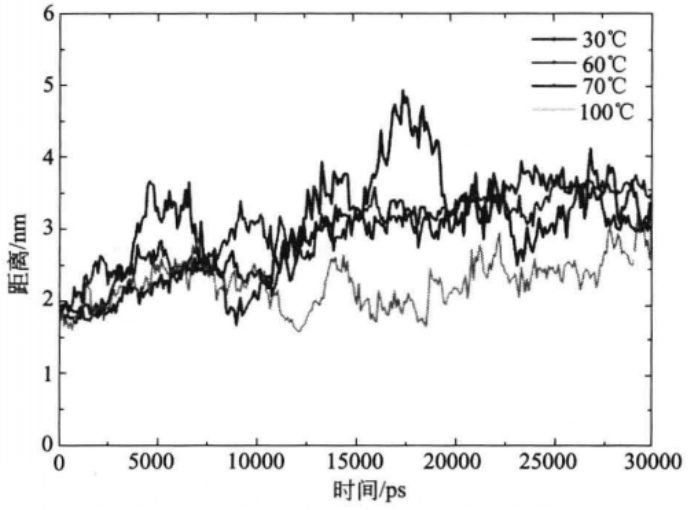



图 5.9　模拟过程中螺旋 6 与螺旋 9 之间的距离（见彩图）

其他温度就会更长。

为了进一步了解 T1 脂肪酶是如何通过温度调节催化活性的分子机理，我们计算整个模拟过程中，各温度下，活性中心附近 5Å 以内的水分子个数[见图 5.11(a)]。通过计算平均值，我们发现在 70℃时，活性中心附近的水分子最多，60℃的标准偏差最小，这就是为什么 70℃时 T1 脂肪酶的活性最高，60℃脂肪酶的活性最稳定。我们再将 70℃时的活性中心附近水分子数，与两个螺旋之间的距离随时间变化值进行比较 [见图 5.11(b)]，发现它们的波动趋势基本一致，证明这两个螺旋组成了 T1 脂肪酶的活性口袋，它们的运动直接影响到 T1 脂肪酶的活性。

同时我们还分析了组成螺旋 6 和螺旋 9 的氨基酸的疏水性，发现螺旋 9

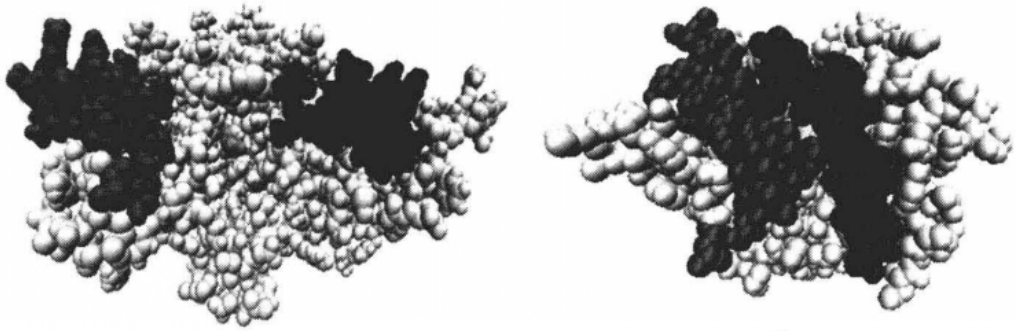



图 5.10　MD 模拟中 T1 脂肪酶打开（a）与关闭（b）的构象（见彩图）
红色代表螺旋 9，蓝色代表螺旋 6，绿色代表活性中心

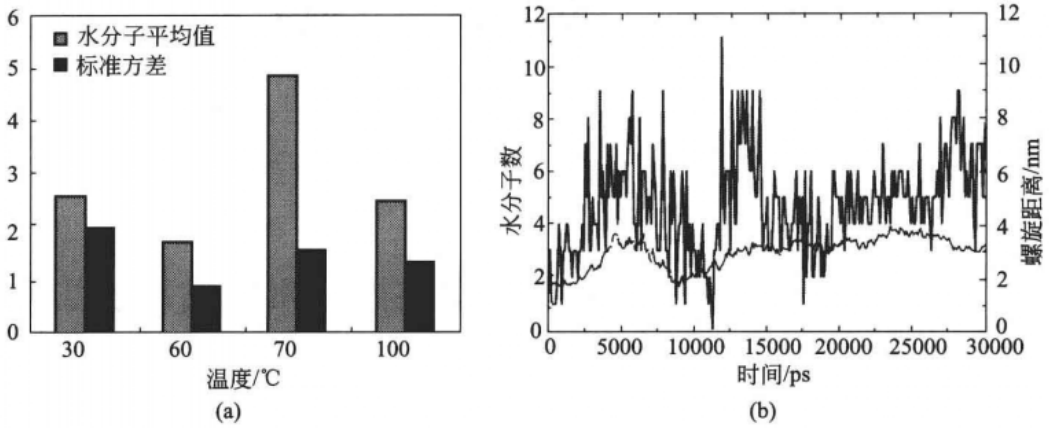



图 5.11　不同温度下活性中心附近水分子统计数据
（a）各温度下活性位点水分子平均值与标准方差；
（b）70℃两个螺旋距离与活性中心水分子数之间的关系

是一个疏水螺旋，而螺旋 6 是一个半疏水螺旋，它靠近活性中心的一半是由疏水氨基酸组成，而靠近溶剂的一面是由亲水氨基酸组成（见图 5.12）。由此我们推测，分子之间的疏水相互作用是导致螺旋运动的主要作用力。在水溶液中，螺旋 6 的亲水部分为脂肪酶提供足够的能量，将活性口袋打开，使底物进入口袋与活性中心结合，催化水解底物。

### 5.5.4　结论

我们通过分子动力学模拟的方法，了解 T1 脂肪酶分子在不同温度条件下的结构变化，从而解释了为什么 T1 脂肪酶在不同温度下的催化活性不同。提出了一种双盖子运动的模型，揭示了 T1 脂肪酶的催化机理，为酶的设计提供了理论基础。

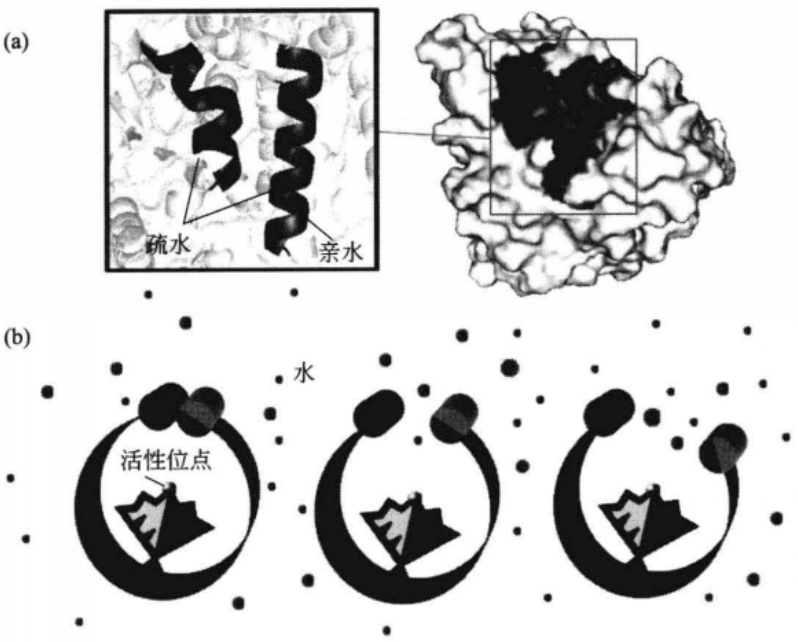



图 5.12　T1 脂肪酶结构变化的分子机理示意图（见彩图）
红色代表疏水，蓝色代表亲水

## 5.6　分子动力学模拟研究氨基酸突变对蛋白结构的影响

利用分子动力学模拟的方法还可以研究蛋白质某一位点的氨基酸突变后，导致蛋白质结构和功能的变化。这里我们以 GTP 结合蛋白 Rab5a 的分子模拟研究为例，了解具体的实验步骤和如何分析模拟得到的分子动力学轨迹数据。

### 5.6.1　研究背景

GTP 结合蛋白可以被称为鸟苷酸结合蛋白或者 G 蛋白。这类蛋白质是一类非常重要的信号转导因子。Rab5a 是 GTP 结合蛋白 Rab 家族的重要一员。这种蛋白质主要是负责调控细胞内吞作用中内涵体的熔合。Rab5a 的空间结构接近球形，主要由 α 螺旋和 β 片层相互折叠围绕而成，如图 5.13 所示。目前已发现 N 端区的主要功能结构域有 P 环、开关区 Ⅰ 和开关区 Ⅱ。Rab5a 的 P 环序列由 27GESAVGKS34 组成，已知此部位突变可引起 Rab5a 的 GTP 水解能力、结合能力和生物活性改变。开关区 Ⅰ 和开关区 Ⅱ 是 Rab5a 蛋白的核苷酸结合位点，分别包括第 49～51、79～81 氨基酸残基。在与 GTP 结合时，两个开关区域形成一个有序而广泛的疏水表面，而在与 GDP 结合时，则表现出一种移位的、活动的开关区域。Rab5a 的中间区域

是与其他 Rab 蛋白高度同源的保守序列，也是 Rab5a 与其辅助蛋白和效应蛋白相互作用的部位，并参与 GDP 和 GTP 的结合。

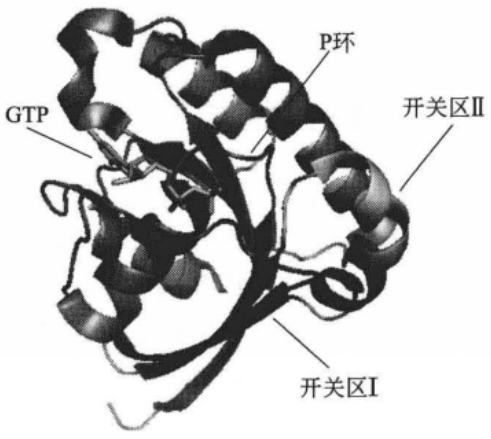



图 5.13　GTP 结合蛋白 Rab5a 的空间结构示意图

## 5.6.2　方法介绍

### 5.6.2.1　初始结构的准备

此研究选取了四个不同的体系来进行 Rab5a 的分子动力学模拟。研究人员首先从蛋白质三级结构数据库 PDB 中提取了 1N6H.pdb 和 1N6L.pdb 两个文件，分别作为野生型 Rab5a 和它的 A30P 突变与 GTP 分子形成的复合物体系。这说明实验的初始结构是通过真实实验获得的分辨率分别为 1.51Å 和 1.60Å 的晶体结构。此外，还将这两个复合物体系中的 GTP 分子删除，从而形成两个新的体系，即野生型 Rab5a 和它的 A30P 突变蛋白体系。然后将晶体结构中的其他杂原子和水分子都删除，最后得到四种体系的初始结构模型。

### 5.6.2.2　模拟方法与参数设置

分子动力学模拟是利用 AMBER8.0 软件来实现的。用 AMBER 软件进行动力学模拟时，一般都要经过分子体系拓扑文件的准备、能量最小化、平衡系统、体系升温和运行真正的动力学模拟等过程，以下是具体操作。首先在模拟中，加入了边界性周期条件和 Amber 的蛋白质力场参数。利用 AMBER8.0 软件中的 t-leap 软件包生成各个体系中的所有原子的拓扑结构、坐标以及添加电荷，并且在开始分子动力学模拟之前，各体系的晶体结构会被放入到充满 TIP3 水分子的模拟盒子中，盒子的边缘距离蛋白质的表面 1.0nm。紧接着，进行能量最小化，直到体系能量的容差达到 100kJ/mol 为止。然后，体系中所有的原子会在 310K 的温度下平衡一段时间。在这段时间内，蛋白质的骨架结构是固定的，以防止其三级结构被破坏。平衡好之后，体系将会被逐渐放松，并且一点点地从 0℃加热到 37℃，最终实现 10ns 的分子动力学模拟。在模拟中，始终保持常温和常压，其耦合常数分别为 0.1ps 和 1.0ps。对于溶剂水分子来说，其等温压缩率设定为 $4.5\times10^{-5}$/bar[1]。此外，体系中的静电势能是利用 PME 算法来计算的；范德华

[1] 1bar=$10^5$Pa。

势能计算中的 CUTOFF 被设定为 12Å。所有的分子动力学模拟的时间步长为 2fs，而体系坐标则是每隔 1ps 记录一次。

### 5.6.3 实验结果与分析

分子动力学模拟轨迹分析。通常来说，分子动力学模拟中只有在体系达到动态平衡之后的构象才具有采集的价值，并且可以被应用于后续的分析中。可以用体系的总能量和均方根偏差（RMSD）来作为衡量是否达到动态平衡的标准，同时，均方根偏差还可以作为衡量体系是否收敛的标准。如图 5.14 所示，四个体系在模拟的过程中，系统的总能量基本上保持平稳。而且在结合了小分子 GTP 之后，野生型和突变型蛋白体系的总能量明显降低，这说明 GTP 分子的结合可以稳定体系。此外，在没有结合 GTP 时，野生型和突变型蛋白体系的总能量相差较大；随着 GTP 分子的结合，这种差距明显减小。同时，如图 5.14 所示，在 10ns 的分子动力学模拟过程中，结合有小分子 GTP 的体系其均方根偏差要略高于没有结合小分子的体系。这说明，小分子的结合可以增强蛋白质的柔韧性。

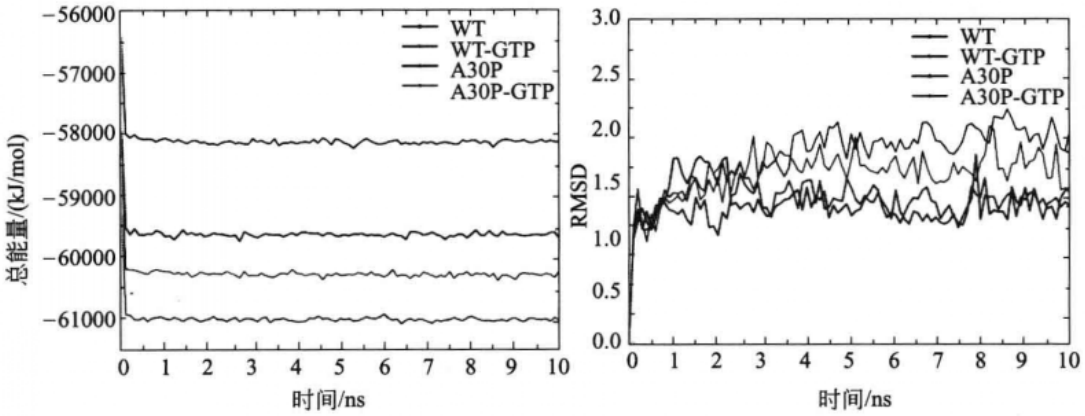



图 5.14 模拟过程中体系总能量和骨架碳原子的均方根偏差随时间变化示意图（见彩图）

再通过观察 10ns 的分子动力学模拟过程中蛋白质骨架碳原子的均方根波动的示意图（图 5.15），可以发现四个体系蛋白质骨架碳原子的均方根波动是相似的，但是在 10-20、30-45、50-60、65-80、110-120 和 140-150 这些氨基酸残基区域，四种体系的骨架碳原子均方根波动有着明显差异。这些区域主要集中在 P 环（Gly27-Ser34）、开关Ⅰ区（Gln49-Ala56）和开关Ⅱ区（Ala77-Leu85）。

在 P 环区域，蛋白质骨架碳原子的均方根偏差（见表 5.2），结合了 GTP 分子的两个体系要明显低于没有结合任何小分子的体系。这说明，GTP 的结合可以降低 P 环区域的结构柔韧性，也就是说 GTP 分子的结合可

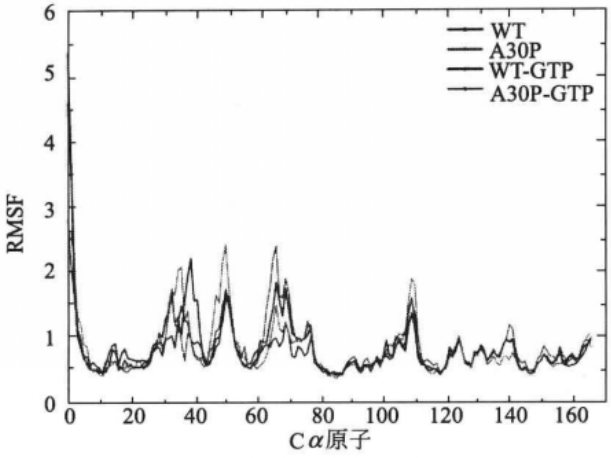



图 5.15　分子动力学过程中骨架碳原子的
均方根波动示意图（见彩图）

以有效减少突变所带来的结构差异。如图 5.16 所示。在 P 环区域由突变所带来的结构变化主要集中在 Ser29 和 Ala30/Pro30 这两个氨基酸残基上。

**表 5.2　四个体系中各个部分的均方根偏差的平均值列表**

| 体 系 名 称 | 蛋 白 整 体 | P 环 | A30/P30 | Res. 49 - 51 |
|---|---|---|---|---|
| WT | 1.30 ± 0.21 | 0.52 ± 0.11 | 0.08 ± 0.02 | 1.23 ± 0.30 |
| A30P | 1.25 ± 0.18 | 0.54 ± 0.10 | 0.04 ± 0.02 | 1.03 ± 0.29 |
| WT-GTP | 1.71 ± 0.31 | 0.36 ± 0.07 | 0.03 ± 0.01 | 0.87 ± 0.21 |
| A30P-GTP | 1.54 ± 0.25 | 0.38 ± 0.08 | 0.03 ± 0.01 | 0.85 ± 0.21 |

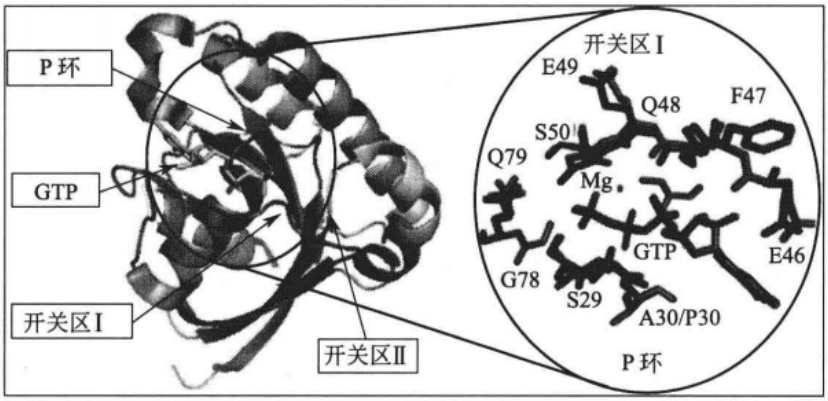



图 5.16　A30P 突变所带来的结构变化示意图（见彩图）
图中野生型结构用蓝色棍状模型表示，而突变型结构则是用红色棍状模型表示

氢键在受体与配体相互结合的过程中发挥着非常重要的作用，研究结合位点上的氨基酸与配体间的距离可以反映受体与配体之间是否发生相互作用。因此通过追踪 30 位置上氨基酸残基骨架氮原子与 GTP 分子的 β-γ 桥连

氧原子之间的距离，如图 5.17 所示，研究者发现 Pro30 的骨架氮原子不能与 GTP 分子的 β-γ 桥连氧原子形成氢键相互作用。而通过 NMR 实验证实，野生型蛋白质中 Ala30 的骨架氮原子与 GTP 分子的 β-γ 桥连氧原子所形成的氢键相互作用在 GTP 结合蛋白质中对于催化 GTP 水解有着重要的作用。同时，30 位氨基酸残基上的骨架氮原子能够稳定催化过程中积累的负电荷。由此可见，第 30 位氨基酸残基的突变可以使 GTP 结合蛋白失去其生物学活性。

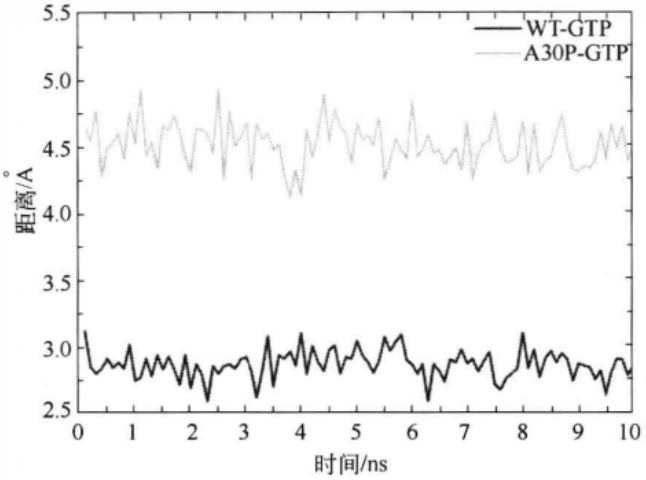



图 5.17　第 30 位氨基酸骨架氮原子与 GTP 分子 β-γ 桥连氧原子之间的距离

同样的均方根偏差变化也发生在开关 I 区域。此外，通过追踪 GTP 分子与周围水分子的距离，发现这些在开关 I 区域的结构变化是由第 30 位突变所带来的 P 环区域和开关 I 区域之间的水分子重排所导致的。在野生型蛋白质体系中，GTP 分子中磷酸根基团上的 3 个氧原子可以分别与 1 个水分子、Thr38 和 Gln65 形成氢键相互作用；而另一个水分子则可以与镁离子形成配位键。但是在 A30P 突变型中，由于 Ser15 和 Lys19 侧链基团的相互接触，使得上述的水分子不能与 GTP 分子形成氢键相互作用。

研究者还定义了第 30 位氨基酸残基与 49 位至 52 位氨基酸残基之间的距离为 GTP 结合蛋白 Rab5a 结合 GTP 口袋的直径。通过测算这一距离，如图 5.18 所示，可以发现，这一水分子重排能够通过其与 49 位至 52 位氨基酸残基以及 GTP 分子的相互作用来实现蛋白质的结构变化，使得 GTP 结合蛋白 Rab5a 结合 GTP 口袋张开，更加有利于 GTP 分子的结合。

### 5.6.4　结论

通过以野生型和 A30P 突变型 GTP 结合蛋白 Rab5a 以及它们与 GTP 分子所形成的复合物为研究对象，通过 10ns 的分子动力学模拟研究 GTP 结合蛋白 Rab5a 的结构波动，以及其 P 环、分子开关 I 区和分子开关 II 区的重

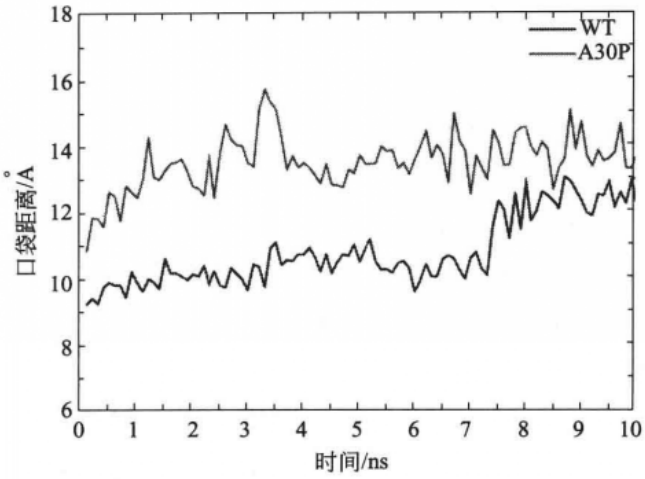



图 5.18　第 30 位氨基酸残基与 49 位至 52 位氨基酸残基之间的距离

要结构变化。研究者发现在 P 环和分子开关Ⅰ区，GTP 分子的结合可以缩小甚至消除第 30 位氨基酸突变所带来的结构变化，使两者的结构波动趋于相似。GTP 分子的这一作用是通过诱导该区域水分子重排来实现的。此外，由于在第 30 位进行了突变，使得突变型蛋白的第 30 位氨基酸残基不能与 GTP 的 β-γ 桥连氧原子形成氢键。而这个氢键被认为是 GTP 结合蛋白中催化 GTP 水解的关键所在，如果失去这个氢键，将导致蛋白质生物学活性的消失。

通过这个实例让我们了解，如何运用分子动力学模拟方法，研究蛋白关键位点的氨基酸突变对蛋白结构和功能的影响，而且证明了分子动力学模拟方法对于解决此类问题的有效性。

## 5.7　靶标蛋白与抑制剂的相互作用研究

分子动力学模拟方法不仅可以用于研究单独的蛋白质分子，而且还可以研究蛋白质与小分子之间的相互作用，了解其分子机制，为实际操作提供理论依据。现在许多科学家将分子动力学模拟运用到药物设计方面，研究蛋白酶与其抑制剂之间的相互作用，找出病毒产生耐药性的原因，为设计更高效的抑制剂提供帮助。所以我们以 HIV 耐药性分子模拟为例，看看分子动力学模拟到底给药物设计提供了怎样的帮助。

### 5.7.1　研究背景

艾滋病的医学全名为“获得性免疫缺陷综合征”，它是由人体免疫缺陷病毒（HIV）引起的。HIV 属于逆转录病毒系慢病毒亚科，潜伏期较长，免疫系统的细胞为主要感染对象，如 T 细胞和巨噬细胞。HIV 有Ⅰ型和Ⅱ

型两种，两者又分别有许多亚型。HIV-Ⅰ的致病力强，是引起全球艾滋病流行的主要病源。随着对 HIV 病毒学及分子生物学等方面的研究，人们逐渐认识到 HIV 蛋白酶是一种对于该病毒复制周期正常运转和病毒粒的成熟、变得有感染能力至关重要的酶。抑制了 HIV 蛋白酶的活性或将此酶的活性降低到极低水平，HIV 在被感染的细胞中就会产生不成熟的、没有感染性的病毒颗粒，从而使得病毒不能正常装配。因此，HIV-Ⅰ蛋白酶是病毒成熟所不可或缺的，这使其成为抗 HIV/AIDS 药物的一个重要靶酶。

HIV-Ⅰ蛋白酶属于天冬氨酰基蛋白酶家族，它有天冬氨酰基蛋白酶共有的 DTG 活性序列。HIV 蛋白酶的三维结构显示出它是由两条含 99 个氨基酸的多肽链形成的 C2 对称的均二聚体。活性部位是由 Asp25/25′-Thr26/26′-Gly27/27′组成的，Asp25/25′在催化中起着关键的作用。HIV-Ⅰ蛋白酶的活性位点位于一个狭长的“通道”的底部，当与抑制剂结合以后，蛋白酶的结构，尤其是挡板区（flap）的结构将发生很大的变化。通道位于二聚体之间的接触面上，通道的长度约为 2.3nm。HIV-Ⅰ蛋白酶具有二重对称性，对称轴位于两个催化天冬氨酸的中间。从两个天冬氨酸的中间开始，两侧分别排列着 S1、S2、S3 和 S1′、S2′、S3′亚位点。

HIV 典型遗传特征之一是高度遗传变异性。据报道，在 HIV-Ⅰ蛋白酶单个亚基的 99 个氨基酸位点中，至少有 49 个位点发生了超过 87 种突变，这其中很多位点的突变已被证实与 HIV 病毒对一种或多种蛋白酶抑制剂产生抗性相关。因此，我们所介绍的研究采用了 B 亚型野生型（*Bwt*）、F 亚型野生型（*Fwt*）以及 2 种来自巴西抗艾滋病失败疗法病患体内的 HIV-Ⅰ蛋白酶突变体 B 亚型突变型（*Bmut*）、F 亚型突变型（*Fmut*）四种蛋白酶与抑制剂 TL-3 分子的晶体结构为研究对象，其对应的 PDB ID code 分别为：*Bwt*—2P3B，*Bmut*—2P3A，*Fwt*—2P3C，*Fmut*—2P3D。通过比较它们结构上的差异，来寻找产生耐药性的根源。

### 5.7.2 实验方法

分子动力学模拟以及模拟前的能量最小化都是使用 AMBER8.0 模拟软件包。各蛋白酶的力场参数使用的是 parm99 力场，而小分子抑制剂的力场通过 AMBER8.0 中的 Antechamber 程序进行构建。初始坐标提取自 PDB 晶体结构，H 原子通过 AMBER8.0 的 LEAP 模块加入。在蛋白酶与抑制剂复合物周围加有 6 个氯离子以保持体系的电中性。使用 SHAKE 算法进行键约束计算（包括 H 原子）。将复合物质心置于截顶八面体盒子中心，选择溶质原子到盒子壁的距离至少为 10Å，然后加入水分子，水分子采用 TIP3P 模型。接着，采用 AMBER8.0 的 PMEMD 模块，以最陡下降法对复合物进行了 1000 步的体系能量最小化。然后将体系在 300K 温度下平衡 20ps。最

后在 298K 下进行了 10ns 的分子动力学模拟。运动方程积分步长（时间步长）取 2fs，每隔 1ps 记录一次坐标文件，以用于后续的分析。在分子动力学模拟中，长程静电相互作用采用 Particle Mesh Ewald（PME）算法。

氢键及疏水相互作用的定义：如果两个重原子为氢的供体和受体，那么它们之间的距离在 3.5Å 之内就被认为是有可能存在氢键相互作用；如果蛋白质中的某个疏水基团距离抑制剂分子 6.5Å 之内，就认为它们之间存在某种疏水相互作用。

### 5.7.3 体系分子动力学轨迹分析

抑制剂 TL-3 与各种蛋白酶的复合物体系在平衡 20ps 后分别进行了 10ns 的分子动力学模拟。通过观察各体系的 RMSD 值（见图 5.19）发现，四个体系的 RMSD 的变化范围小于 3.0Å，各体系在模拟 4.0ns 后都达到了动态平衡。然后进行蛋白酶各氨基酸骨架 Cα 原子在整个 10ns 分子动力学模拟中相对于初始输入结构的综合振动计算，以了解 TL-3 抑制剂分子的结合对 HIV-Ⅰ蛋白酶稳定性的影响，结果见图 5.20。从图 5.20 可以看出，四个体系的蛋白酶 A 链（氨基酸残基 1～99）的振动情况（RMSF）均不同于 B 链，暗示着对称性抑制剂分子 TL-3 与蛋白酶采用的是不对称结合方式。

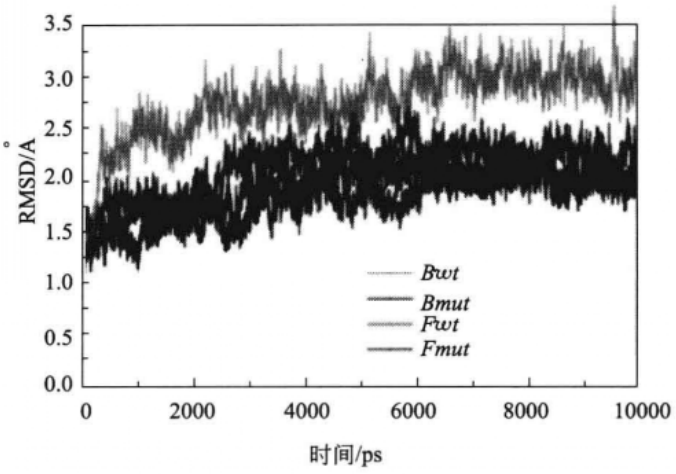



图 5.19 各体系蛋白酶骨架原子的 RMSD 随时间变化规律（见彩图）

氢键和疏水相互作用：非键相互作用对于维持蛋白结构、识别过程、蛋白质-配体相互作用都是非常重要的。其中的氢键和疏水相互作用在抑制剂与蛋白质之间的亲和力大小发挥着至关重要的作用。因此，此研究进行了定量分析和评价这些氢键，计算了 10ns 分子动力学模拟中所有氢键的存活时间（或者称为“寿命”和“Lifetime”），如表 5.3 所示。*Fmut* 蛋白酶与 TL-3 抑制剂之间没有一个氢键的寿命大于 12%，也就是说它们之间几乎没有氢键稳定形成。对于 *Bwt* 和 *Bmut* 蛋白酶，氢键主要形成于挡板区（flap）

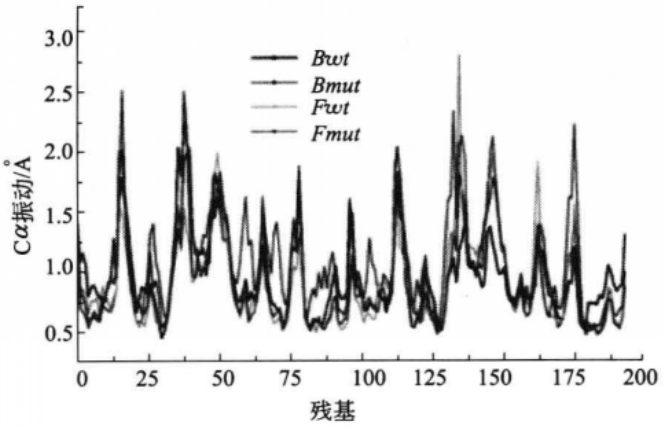



图 5.20　各体系中骨架碳原子在分子动力学模拟中的振动情况（见彩图）

[如表格中的 Ile50（B）、Gly48、Gly48（B）、Gly49（B）] 和催化三角区（如表格中的 Asp25、Asp29）。而对于 Fwt 蛋白酶，氢键主要形成于催化三角区 [如表格中的 Asp25，Gly27（B），Asp29（B），Asp30（B）]。分子动力学模拟所得到的这个结果与实验结果总体趋势是一致的，即耐药性随着蛋白酶与抑制剂分子之间氢键相互作用的减少而上升，如表 5.4 所示。

**表 5.3　各个蛋白酶与抑制剂分子之间的氢键存活时间**

| 项目 | 氢键供体 | 氢键受体 | *Bwt* | *Bmut* | *Fwt* | *Fmut* |
|---|---|---|---|---|---|---|
| 挡板区（flap） | Ile50(B)N | TL-3 O52 | 77.57% | 60.45% | — | — |
| | Gly48N | TL-3 O8 | 59.72% | 21.29% | 18.68% | — |
| | Gly49(B)N | TL-3 O52 | — | 50.51% | — | — |
| | TL-3 N54 | Gly48(B)O | 43.08% | — | — | — |
| | TL-3 N52 | Gly48(B)O | 13.51% | 67.13% | — | — |
| | Gly48(B)N | TL-3 N54 | 34.10% | — | — | — |
| | Ile50N | TL-3 O52 | — | 36.02% | — | 11.69% |
| 催化区（catalytic） | TL-3 O1 | Asp25OD1 | 49.07% | — | — | — |
| | TL-3 O1 | Asp25OD2 | 36.17% | — | 74.16% | — |
| | TL-3 O51 | Asp25OD1 | 26.00% | — | — | — |
| | TL-3 O51 | Asp25OD2 | 16.95% | — | — | — |
| | TL-3 O1 | Asp25(B)OD1 | 24.52% | — | — | — |
| | TL-3 O1 | Asp25(B)OD2 | 12.67% | — | 14.13% | — |
| | TL-3 N52 | Gly27(B)O | — | — | 77.87% | — |
| | Asp29N | TL-3 O4 | — | 42.26% | — | — |
| | Asp29(B)N | TL-3 O59 | — | — | 73.51% | — |
| | Asp30(B)N | TL-3 O59 | — | — | 57.00% | — |
| | TL-3 N54 | Asp29(B)OD2 | — | — | 28.58% | — |
| 共计 | | | 3.93 | 2.77 | 3.44 | 0.12 |

注：只有那些存活时间超过 10%的氢键相互作用才被在表中列出，挡板区包括 48 位至 58 位氨基酸，催化区包括 23 位至 32 位氨基酸。

表 5.4　蛋白酶与抑制剂分子氢键相互作用与耐药性之间的关系

| 项目 | *Bwt* | *Bmut* | *Fwt* | *Fmut* |
|---|---|---|---|---|
| 氢键的平均数目 | 3.93 | 2.77 | 3.44 | 0.12 |
| $K_i$ 值 | 1 | 3 | 7 | 54 |

注：耐药性以实验测得的抑制常数 $K_i$ 为 *Bwt* 蛋白酶的多少倍来表示。

另一方面，针对上市的口服药物的市场调研结果显示，平均每个药物包含 1～2 个氢键供体和 3～4 个氢键受体，而一个药物中疏水原子的平均数目有 16 个之多。这些数据提示我们在药物诱导结合到受体上的过程中，疏水相互作用是占主导地位的驱动力。有关 10ns 分子动力学模拟中疏水作用的数据统计见表 5.5。从表格中可看出，*Bmut* 蛋白酶与 TL-3 抑制剂分子之间的疏水作用与 *Bwt* 体系大致相同。与 *Bwt* 体系相比，*Fwt* 蛋白酶与抑制剂之间在 S1/S1′疏水口袋少了 3 个疏水作用；*Fmut* 蛋白酶与抑制剂之间在催化三角区则少了 6 个疏水作用。这些减少消失的疏水相互作用很可能就是引起 *Fwt*、*Fmut* 蛋白酶耐药性的主要原因。

表 5.5　分子动力学模拟中各蛋白酶疏水相互作用统计

| 项目 | *Bwt* | *Bmut* | *Fwt* | *Fmut* |
|---|---|---|---|---|
| 催化区 | 7 | 6 | 9 | 1 |
| 挡板区 | 4 | 4 | 4 | 4 |
| S1/S1′口袋 | 6 | 6 | 3 | 9 |
| 其他区域 | 0 | 1 | 2 | 1 |
| 共计 | 17 | 17 | 18 | 15 |

注：表中所表示出的是各个蛋白酶与抑制剂分子之间存活时间大于 50%的疏水相互作用个数。

## 5.7.4　结构变化分析

### 5.7.4.1　*Bmut* 和 *Bwt* 蛋白酶之间的结构差异

*Bmut* 体系与 *Bwt* 体系相比，主要差异在于氢键而不是疏水相互作用。表 5.3 显示，在 *Bmut* 蛋白酶与抑制剂的复合物中，抑制剂 TL-3 中心二醇的羟基与蛋白酶催化活性中心的 Asp25/Asp25（B）的氧原子 OD1/OD2 之间关键性的氢键消失了，取而代之的却是 Asp29 的 N 原子与抑制剂分子的羰基原子 O4 之间的氢键。之所以说抑制剂中心二醇的羟基与蛋白酶催化中心的 Asp25/Asp25（B）之间形成的氢键非常关键，主要是因为在两个催化天冬氨酸［Asp25/Asp25（B）］之间只能容纳、结合（通过氢键的方式）抑制剂二醇中心的一个羟基，学者们认为这种不对称结合仿效了天冬氨酰蛋白酶催化水解反应中的那个四面体型过渡态，有效地占据了蛋白酶天然底物的结合位点，从而实现了竞争性抑制的抑制剂疗效。因此，这么关键的氢键的消失很可能就是降低抑制剂分子 TL-3 与 *Bmut* 蛋白酶之间结合亲和力的

根本因素，从而最终导致 *Bmut* 蛋白酶耐药性的产生。

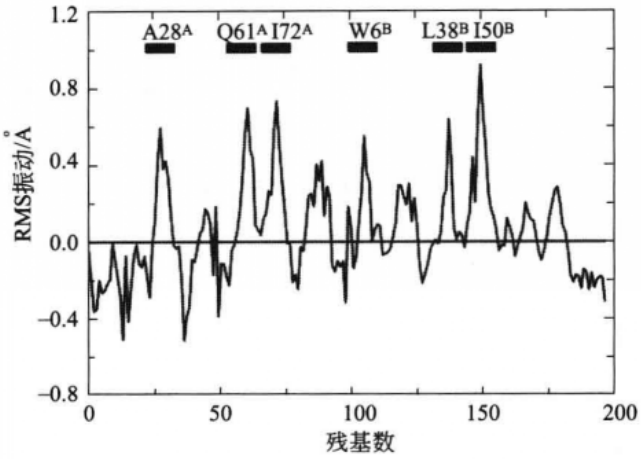



图 5.21　分子动力学模拟过程中 *Bwt* 和 *Bmut* 蛋白酶 RMSF 差异

图 5.22　*Bmut* 蛋白酶（红色）与 *Bwt*（黄色）蛋白酶的空间结构叠合（见彩图）

图 5.21 显示，*Bmut* 蛋白酶的氨基酸位点 Ala28、Gln61、Ile72、Trp6（B）、Leu38（B）、Ile50（B）相对于 *Bwt* 蛋白酶振动幅度要大。对氢键作用影响最大的就是催化三角区的 Ala28 的大幅度偏移，使得抑制剂二醇中心的羟基与 Asp25/Asp25（B）之间的重要氢键消失了。从 *Bwt* 与 *Bmut* 蛋白酶的平均构象（与计算出的平均结构最接近的时刻点构象）的叠合比对图（见图 5.22）中可以发现，活性位点氨基酸中偏移最大的就是 Asp25 的骨架 Cα 原子——*Bmut* 的 Asp25 向远离结合腔的方向偏移了 3.32Å。为什么会有这样的位置偏移呢？原来，*Bmut* 蛋白酶发生了 L90M 突变，90 位的氨基酸 Met 的侧链比 *Bwt* 蛋白酶的 Leu 变得体积更大，使得 Met90 与 *Bmut* 蛋白酶的催化环（残基 24～26）形成了更多的范德华相互作用。因此，推测 *Bmut* 蛋白酶氨基酸之间增加的范德华相互作用限制了氨基酸残基 24～29 的骨架的自由运动能力，大大降低了氨基酸残基 24～29 的骨架的灵活性，

因而当抑制剂分子 TL-3 结合到 *Bmut* 蛋白酶上时，这些活性位点的肽链很“僵硬”，不能做出相应的结构调整去更好地与 TL-3 接触、结合。

#### 5.7.4.2 *Fwt* 和 *Bwt* 蛋白酶之间的结构差异

*Fwt* 体系与 *Bwt* 体系相比，*Fwt* 蛋白酶与 TL-3 之间的氢键主要形成于催化三角区的 Asp25、Gly27（B）、Asp29（B）、Asp30（B），而 *Bwt* 蛋白酶除了在催化三角区，还在挡板区形成氢键的 Ile50（B）、Gly48、Gly48（B）。*Fwt* 蛋白酶在挡板区缺失的氢键与其 B 链挡板区较大幅度的振动相一致（见图 5.23）。*Fwt* 蛋白酶的挡板区与抑制剂分子没有形成氢键，这在一定程度上加剧了其耐药性的产生。从 *Fwt* 与 *Bwt* 蛋白酶的平均构象的叠合比对图局部图（图 5.24）中可以看出，相对于 *Bwt* 蛋白酶，包含 M36I 突变的 *Fwt* 蛋白酶的环区主链（残基 33～42）向结合腔方向折叠了稍许。仔细分析突变位点第 36 位氨基酸后发现，突变成较短的 Ile 后，为了维持第 36 位氨基酸位点与周围氨基酸之间的相互作用不变，较短的异亮氨酸的侧链将环区（残基 33～42）的主链向结合腔的方向下拉，所以铰链区的下拉运动带动了挡板区的大幅度振动。

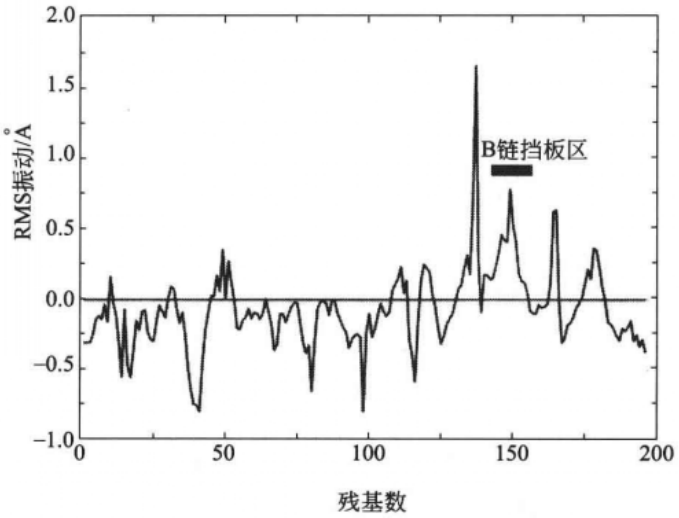



图 5.23 分子动力学模拟中 *Fwt* 和 *Bwt* 蛋白酶 RMSF 差异

此外，*Fwt* 体系与 *Bwt* 体系更大的区别在于疏水作用上的差异。从疏水作用的总数上看，两个体系基本相同，但是从具体的分布上看，差异巨大：相对于 *Bwt* 体系，*Fwt* 蛋白酶催化三角区形成的疏水作用增加了 2 个，S1/S1′疏水口袋形成的疏水作用减少了 3 个。在 *Fwt* 体系中抑制剂 TL-3 的 P1 基团（结构见图 5.25）与蛋白酶的 S1/S1′疏水口袋［Ile50，Pro81（B），Val82（B），Ile84（B）］之间已经不再形成疏水相互作用。通过叠合比对 *Fwt* 与 *Bwt* 体系的平均构象（见图 5.24）发现：*Bwt* 体系中抑制剂 TL-3

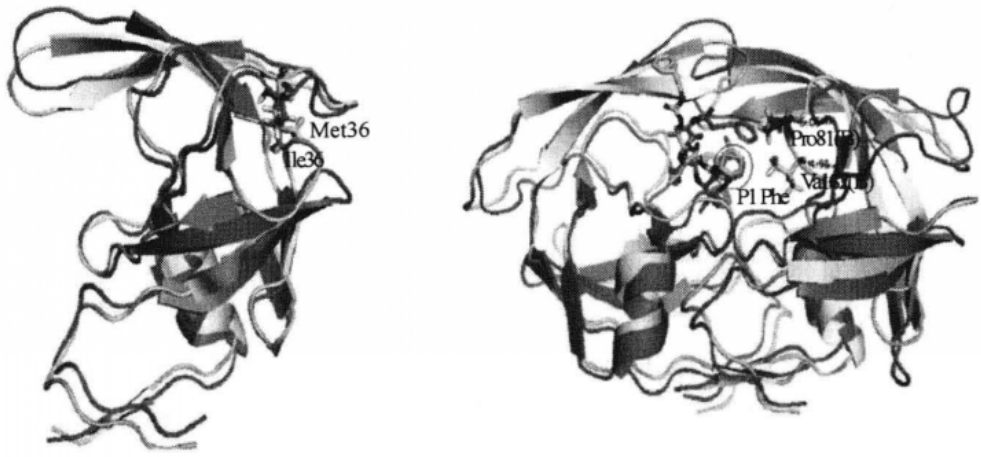



图 5.24　*Fwt*（绿色）和 *Bwt*（黄色）蛋白酶 B 链（a）和 S1/S1′疏水口袋（b）的空间结构叠合（见彩图）

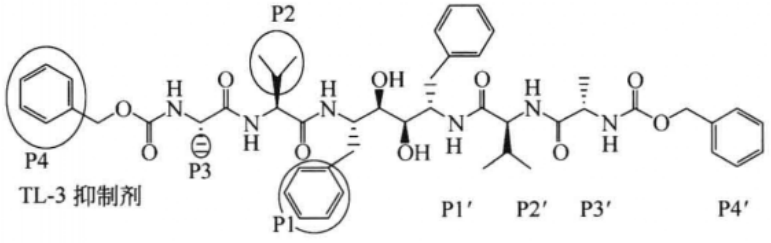



图 5.25　C2 对称的二醇抑制剂分子 TL-3 的化学结构

的 P1 苯环被紧凑地塞在由 *Bwt* 蛋白酶 B 链的“80s”环区的 Pro81（B）、Val82（B）等氨基酸侧链所形成的 S1/S1′疏水口袋中。而在 *Fwt* 体系中，由于 *Fwt* 蛋白酶 B 链的“80s”环区有了很大的偏移，Pro81（B）的 Cα 原子、Val82（B）的 Cα 原子比 *Bwt* 蛋白酶的相应位点向远离结合腔的方向分别偏移了 6.05Å、4.91Å，无法再形成疏水作用，从而降低了抑制剂 TL-3 与 *Fwt* 蛋白酶的结合亲和力，最终导致耐药性的产生。

#### 5.7.4.3　*Fmut* 蛋白酶和 *Bwt* 蛋白酶之间的结构差异

*Fmut* 蛋白酶与 TL-3 抑制剂分子之间没有形成氢键（见表 5.3），形成的疏水作用也较少，尤其是催化三角区只形成了 1 个疏水相互作用（见表 5.5），所以 *Fmut* 蛋白酶呈现出最显著的耐药性，这与实验中测得的抑制常数这个体系最大是相一致的。从 *Fmut* 蛋白酶结构的角度分析，突变 M36I 改变了蛋白酶挡板区的构象，并且突变 V82A 使得抑制剂分子 TL-3 结构重排。因挡板区与 Pro81 的折拢密封了原有的裂缝，抑制剂的 P1′基团的苯环只能离开原来的位置，朝着第 82 位丙氨酸周围的空腔转动（见图5.26）。抑制剂分子这样的结构重排造成 P1′基团与整个 S1′疏水口袋之间的相互作用数目（除了疏水作用，还有其他非键相互作用，如范德华相互作用等）大大

大大减少，再加上新生成的这个空腔使得原来的S1′疏水口袋变得更深，从而导致抑制剂分子与蛋白酶结合得很松散，其结合亲和力大大下降，最终导致 *Fmut* 蛋白酶的耐药性的产生。

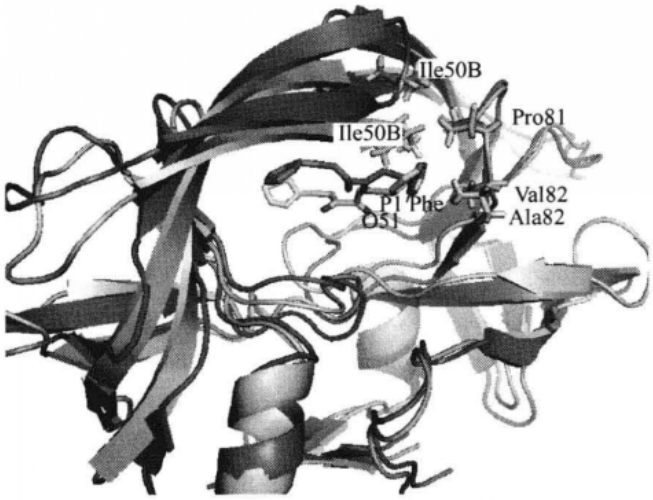



图 5.26　*Fmut*（蓝色）和 *Bwt*（黄色）蛋白酶空间结构叠合（见彩图）

### 5.7.5　结论

HIV-Ⅰ病毒的耐药性是由病毒在其复制周期中发生的突变所造成的，究其根本，则是因为病毒的复制极其迅速，同时其逆转录酶没有校正功能。本实例以四种不同的 HIV-I蛋白酶（*Bwt*、*Bmut*、*Fwt* 和 *Fmut*）与同一种蛋白酶抑制剂分子 TL-3 的复合物晶体为研究对象，运用分子动力学模拟的技术手段，探索组成 HIV-I蛋白酶亚基的 99 个氨基酸位点中哪些位点的突变与HIV-I病毒耐药性的产生相关，又是通过怎样的结构变化导致了抑制剂分子与 HIV-I蛋白酶之间的结合亲和力下降，从而最终引发耐药性的形成。

了解了耐药性的形成原因，我们就可以针对这一结构变化，设计更有效的抑制剂来治疗新型的 HIV 病毒导致的艾滋病，为新药的研发提供理论指导，减少药物开发的前期投入，降低生产成本。

## 5.8　本章总结

前面介绍的几个例子仅仅是比较有代表性的实例，分子模拟在蛋白研究中的应用是非常广泛的，方法也是多种多样的，在这里就不一一介绍了。相信通过以上的例子大家对于分子模拟在蛋白质研究中的应用也初步有些直观的认识，希望大家在以后的学习和工作过程中多多实践，并推动分子模拟应用于蛋白质研究的发展。虽然分子模拟技术日渐成熟，但是目前分子模拟在

蛋白质研究中所起的作用不是决定性的，它更多的是提供给人们结构方面的信息，所以必须与实验相互结合，才能更好地达到研究目的。

## 参考文献


[1] Grant G H. Computational chemistry. Oxford: Oxford University Press, 1995.

[2] Rouvray D. Atoms as hard spheres. Chemistry in Britain, 2000, 36: 25.

[3] Tetko I V, et al. Virtual computational chemistry laboratory - design and description. Journal of Computer-Aided Molecular Design, 2005, 19: 453-463.

[4] Schleyer P R, A N L, Clark T, Gasteiger J, Kollman P A, Schaefer Ⅲ H F and Schreiner P R. Encyclopedia of Computational Chemistry (ed. Sons, J. W.). American Chemical Society, Chichester, U. K, 1998.

[5] Arnold K, Bordoli L, Kopp J & Schwede T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics, 2006, 22: 195-201.

[6] Kopp J & Schwede T. The SWISS-MODEL Repository: new features and functionalities. Nucleic Acids Res, 2006, 34: D315-8.

[7] Schwede T, Kopp J, Guex N & Peitsch M C. SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Res, 2003, 31: 3381-3385.

[8] Eswar N, et al. Tools for comparative protein structure modeling and analysis. Nucleic Acids Res 2003, 31: 3375-3380.

[9] Uchoa H B, et al. Parmodel: a web server for automated comparative modeling of proteins. Biochem Biophys Res Commun, 2004, 325: 1481-1486.

[10] Dalton J A & Jackson R M. An evaluation of automated homology modelling methods at low target template sequence similarity. Bioinformatics, 2007, 23: 1901-1908.

[11] Pieper U, et al. MODBASE, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Res, 2004, 32: D217-222.

[12] Pieper U, et al. MODBASE: a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res, 2006, 34: D291-295.

[13] Pieper U, Eswar N, Stuart A C, Ilyin V A & Sali A. MODBASE, a database of annotated comparative protein structure models. Nucleic Acids Res, 2002, 30: 255-259.

[14] Sanchez R, et al. MODBASE, a database of annotated comparative protein structure models. Nucleic Acids Res, 2000, 28: 250-253.

[15] Bates P A, Kelley L A, MacCallum R M & Sternberg M J. Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. Proteins Suppl, 2001, 5: 39-46.

[16] Montgomerie S, et al. PROTEUS2: a web server for comprehensive protein structure prediction and structure-based annotation. Nucleic Acids Res, 2008, 36: W202-209.

[17] Zhang Y. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics, 2008, 9: 40.

[18] Eisenberg D, Luthy R & Bowie J U. VERIFY3D: assessment of protein models with three-dimensional profiles. Methods Enzymol, 1997, 277: 396-404.

[19] Laskowski R A, Rullmannn J A, MacArthur, M W, Kaptein R & Thornton J M AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J

Biomol NMR, 1996, 8: 477-486.

[20] Wei D Q, Du Q S, Sun H & Chou K C. Insights from modeling the 3D structure of H5N1 influenza virus neuraminidase and its binding interactions with ligands. Biochem Biophys Res Commun, 2006, 344: 1048-1055.

[21] Wang Y, Wei D Q & Wang J F. Molecular dynamics studies on T1 lipase: insight into a double-flap mechanism. J Chem Inf Model, 2010, 50: 875-878.

[22] Abd Rahman R N, Leow T C, Salleh A B & Basri M. Geobacillus zalihae sp. nov., a thermophilic lipolytic bacterium isolated from palm oil mill effluent in Malaysia. BMC Microbiol, 2007, 7: 77.

[23] Matsumura H, et al. Novel cation-pi interaction revealed by crystal structure of thermoalkalophilic lipase. Proteins, 2008, 70: 592-598.

[24] Wang J F & Chou K C. Insight into the molecular switch mechanism of human Rab5a from molecular dynamics simulations. Biochemical and Biophysical Research Communications, 2009, 390: 608-612.

[25] Neitzel K L & Hepler J R. Cellular mechanisms that determine selective RGS protein regulation of G protein-coupled receptor signaling. Seminars in Cell & Developmental Biology, 2006, 17: 383-389.

[26] Ali B R, Wasmeier C, Lamoreux L, Strom M & Seabra M C. Multiple regions contribute to membrane targeting of Rab GTPases. Journal of Cell Science, 2004, 117: 6401-6412.

[27] Gu H, Chen H F, Wei D Q & Wang J F. Molecular dynamics simulations exploring drug resistance in HIV-1 proteases. Chinese Science Bulletin, 2010, 55: 2677-2683.

[28] Perelson A S, Neumann A U, Markowitz M, Leonard J M & Ho D D. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. Science, 1996, 271: 1582-1586.

[29] Earl P L, Doms R W & Moss B. Oligomeric structure of the human immunodeficiency virus type 1 envelope glycoprotein. Proc Natl Acad Sci U S A, 1990, 87: 648-652.

[30] Weiss C D, Levy J A & White J M. Oligomeric organization of gp120 on infectious human immunodeficiency virus type 1 particles. J Virol, 1990, 64: 5674-5677.

[31] Appelt K. Crystal structures of HIV-1 protease-inhibitor complexes. Perspectives in drug discovery and design, 1993, 1: 23.

[32] Haase A T. Population biology of HIV-1 infection: viral and $CD4^+$ T cell demographics and dynamics in lymphatic tissues. Annu Rev Immunol, 1999, 17: 625-656.

[33] Ohtaka, H. & Freire, E. Adaptive inhibitors of the HIV-1 protease. Prog Biophys Mol Biol, 2005, 88: 193-208.

[34] 曾庆平. 人类艾滋病. 北京: 人民卫生出版社: 2001.

[35] Caride E, et al. Genotypic and phenotypic evidence of different drug-resistance mutation patterns between B and non-B subtype isolates of human immunodeficiency virus type 1 found in Brazilian patients failing HAART. Virus Genes, 2001, 23: 193-202.

# 第6章 锌指核酸酶的构建

程超　李荣秀

锌指蛋白（zinc finger protein）是包含一系列锌指结构的DNA特异性结合蛋白质，是转录因子的一种，与基因表达、基因调控、细胞分化和胚胎发育等密切相关。锌指核酸酶（zinc finger nuclease）是由锌指蛋白和非特异性核酸内切酶构成，发挥特异性核酸内切酶的作用，其中锌指蛋白起到识别并结合特异碱基序列的作用。锌指核酸酶的出现，不仅大大提高了基因组靶向修饰的特异性，而且效率也提高了，已报道的转化率达到29%，甚至更高。锌指核酸酶技术给重组人类基因组带来了希望，使基础研究具有基因治疗的意义。本章介绍了锌指蛋白的结构和功能，并以镰刀形红细胞贫血症为例进行实验设计，构建锌指核酸酶用于基因治疗。

马萨诸塞综合医院病理学家、哈佛大学副教授J. Keith Joung是锌指协会（http：//www. zincfingers. org）的主要领导者，他带领八个实验室组成的协会提出了制造“锌指核酸酶”的一种开源方式。明尼苏达大学基因工程中心主任Daniel Voytas也是锌指协会的创办者之一，他的实验室利用这种方法创造的锌指核酸酶对烟草细胞进行修改，从而使其具备了抗除草剂的特性。

## 6.1 锌指蛋白

锌指蛋白是一种蛋白质结构基元，常出现在DNA结合蛋白中，一个含有大约30个氨基酸的环和$Zn^{2+}$形成手指状的结构，$Zn^{2+}$与环上的4个半胱氨酸（Cys）或2个半胱氨酸和2个组氨酸（His）形成配位。锌指蛋白通常由一系列锌指组成，每一个锌指单位指部可以伸入DNA双螺旋的深沟，接触5个核苷酸。锌指蛋白最初在非洲爪蟾中发现，在动物、植物以及微生物中都有广泛的分布。

### 6.1.1 锌指蛋白的分类

蛋白质的结构与功能密切相关，根据锌指蛋白保守结构中半胱氨酸（C）和组氨酸（H）的数目和位置的不同，可将锌指蛋白分为 $C_2H_2$ 型、$C_2HC$ 型、$(C_2)_2$ 型、$C_4$ 型和 $C_6$ 型等，$C_2H_2$ 型锌指蛋白最为普遍。

#### 6.1.1.1 $C_2H_2$ 型锌指蛋白

$C_2H_2$ 型锌指蛋白也是经典型锌指蛋白，最先在爪蟾 TFⅢA 中发现，是目前研究最多的一类锌指蛋白。这类锌指蛋白能形成简单的 ββα 模型结构，其中 α 链的氨基酸残基能识别目标碱基序列，如图 6.1。$C_2H_2$ 型锌指蛋白以单体的形式与核酸发生相互作用。例如人类的 hZF1 蛋白和人类转录因子 SP1 都是这类蛋白。

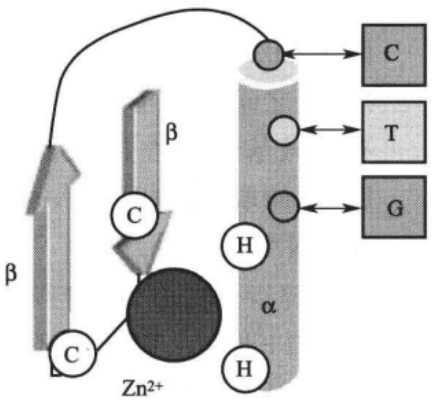


图 6.1 $C_2H_2$ 型锌指蛋白结构示意图

$C_2H_2$ 型锌指蛋白具有靶位点特异性，且其结构简单，因此 $C_2H_2$ 型锌指蛋白被用作人造锌指蛋白的骨架。保持 $C_2H_2$ 型锌指蛋白的基本结构不变，替换识别核酸序列的特定氨基酸残基，再融合其他功能域，就可以得到具有靶向性的人造锌指蛋白。人造锌指蛋白可以抑制或激活靶基因，对 DNA 进行修饰，对靶基因敲除、替换等。因此，人造锌指蛋白应用前景广阔，研究价值显著，是未来人类基因治疗的革命性的工具。

#### 6.1.1.2 $C_4$ 型锌指蛋白

$C_4$ 型锌指蛋白结构中含 8 个保守的 Cys 残基，每 4 个 Cys 残基与一个 $Zn^{2+}$ 配位结合，形成 DNA 识别位点。例如 GATA 蛋白、LIM 蛋白及糖皮质激素等都含有 $C_4$ 型锌指蛋白。$C_4$ 型锌指蛋白以二聚体方式与核酸结合发挥作用，两个亚基的 α-识别螺旋分别与 DNA 大沟相结合，两个 DNA 结合位点间隔约为 3～4bp，通常同源二聚体识别靶基因的反向重复序列，而异源二聚体结合正向重复序列。

#### 6.1.1.3 $C_6$ 型锌指蛋白

$C_6$ 型锌指蛋白由 6 个保守的 Cys 残基与 2 个 $Zn^{2+}$ 以配位键结合，形成 DNA 结合域，以同二聚体形式与 DNA 相结合作用。$C_6$ 型锌指蛋白家族在酵母等真菌中大量存在，如酵母转录因子 Gal4p、Leu3p 等。这类锌指只有一个锌指区域，但是却可以结合 2 个锌离子。

### 6.1.2 锌指蛋白的作用及机制

锌指蛋白依靠其特殊的蛋白质锌指结构，可以和 DNA、RNA 及 DNA-

RNA 杂交双链分子特异性结合，也可以与其他蛋白质发生相互作用。锌指蛋白的这些性质，使其参与细胞分化、增殖和凋亡等多种重要生命过程，对 DNA 识别、RNA 包装、转录及转录后调控、蛋白质折叠和装配、肿瘤的形成和机体免疫等有重要作用。

#### 6.1.2.1 锌指蛋白对核酸的识别

锌指蛋白包含特异性的锌指结构，这种特异性的锌指结构在对 DNA 的特异性识别和结合中起重要作用。锌指蛋白能特异性地识别 DNA，主要是通过锌指的 α 螺旋上的氨基酸与 DNA 双螺旋大槽的特异性接触。小鼠转录因子 Zif268 蛋白与 DNA 结合形成复合物的晶体结构揭示了锌指蛋白与 DNA 之间的相互作用：每个锌指的 α 螺旋与 DNA 上的 3 个碱基结合；所有的碱基特异性接触都位于 DNA 双螺旋的一条链上；每个锌指相对于 DNA 的取向是相同的，与 DNA 成反平行，N 端靠近 DNA 的 3′端，C 端靠近 DNA 的 5′端。

锌指蛋白不仅能与 DNA 发生相互作用，也可以识别 RNA，例如非洲蟾蜍 TFⅢA。TFⅢA 不仅可以和 DNA 结合，也可以和 RNA 结合，但其结合作用位点不同。N 端的锌指 1～3 负责结合 DNA，锌指 4～6 识别结合 5S rRNA。除了一级结构外，磷酸骨架在对 RNA 的识别过程中起到了重要的作用。

锌指蛋白与 DNA-RNA 杂交双链分子也是特异性结合的。实验证明，用人为设计的锌指蛋白分别与 DNA、DNA-RNA 杂交双链分子结合，发现其和 DNA-RNA 杂交双链分子的结合强度是它与双链 DNA 分子结合强度的 5 倍。

#### 6.1.2.2 转录因子

许多锌指蛋白的结构域包含能与核酸特异结合的区域，并与某些效应结构域（如 KRAB、SCAN、BTBPPOZ、SNAG、SANT 和 PLAG 等）相连，这类锌指蛋白常作为转录因子起作用，可调控靶基因的转录。

包含 KRAB 结构域的锌指蛋白是哺乳动物中最大的转录调控因子家族。这种锌指蛋白约占人类基因组中已知锌指蛋白的 1/3，由 57～75 个保守的氨基酸残基组成，可分成 A 盒和 B 盒两个结构域。其中 A 盒在转录抑制中起核心作用，B 盒起辅助作用。SCAN 结构域是一个高度保守的富含亮氨酸的结构域，仅在脊椎动物中存在，起转录抑制作用。实验表明，SCAN 结构域可以调节蛋白质之间的相互作用，并与其他蛋白质协同调节转录。BTB 结构域也称为 POZ 结构域，是一个高度保守的结构域。BTB 结构域由四个 α 螺旋组成，其主要功能是介导蛋白质相互作用形成蛋白质复合体。BTB 蛋白不仅有转录抑制作用，还有转录激活作用，例如 PLZF。$C_4$、$C_6$ 型转录因子也比较常见。$C_4$ 型转录因子，主要是由类固醇受体家族成员组成的。

$C_6$ 型锌指蛋白在酵母等真菌中较为多见。

#### 6.1.2.3 转录后调控

一些锌指蛋白能与 RNA 结合，这类锌指蛋白的功能是参与 mRNA 的成熟剪接以及降解等转录后调控过程。例如锌指蛋白 S27 可以结合损伤的 mRNA 分子，参与后者的加工处理。锌指蛋白与 RNA 相互作用的现象越来越受到关注，但是二者的作用模式尚未研究透彻。

#### 6.1.2.4 介导蛋白质之间相互作用

一些锌指蛋白特异地介导蛋白质之间相互作用，这类锌指蛋白主要包括 LIM 锌指、MYND 锌指、PHD 锌指和 RING 锌指等。例如 LIM 结构域，它是一种介导蛋白质-蛋白质相互作用的锌指基元，是很多 LIM 蛋白行使其生物学功能的关键结构域。LIM 结构域由两个串联的锌指结构组成，富含半胱氨酸和组氨酸。LIM 结构域不与 DNA 结合，而是介导蛋白质间的相互作用，例如可以与转录因子结合，也可以和不能与 DNA 结合的辅助因子结合。对于蛋白质和蛋白质之间相互作用的研究，有助于揭示这些蛋白质的生物学机制。

### 6.1.3 锌指蛋白与免疫性疾病

锌指蛋白在生物体的免疫系统中起到了非常重要的作用，它能在免疫系统的各个环节中起调控作用，如调控 T 细胞、B 细胞的活化及细胞因子的产生，并和临床许多疾病相关。随着研究的深入及人造锌指蛋白的合成，为免疫系统的调节及免疫性疾病的治愈提供了新的解决途径。

#### 6.1.3.1 锌指蛋白与自身免疫病 APECED

自身免疫病 APECED（autoimmune polyendocrinopathycandidiasis-ectodermal dystrophy）又称自身免疫性多腺体综合征Ⅰ（autoimmune polyglandular syndrome type Ⅰ，APS Ⅰ），是一种罕见的常染色体隐性遗传病，主要表现为内分泌器官特异性自身免疫病，通常还伴有白色念珠菌感染，是唯一一个单基因缺乏导致的免疫性疾病。自身免疫调节因子（autoimmune regulator，AIRE）是一种具有转录活化潜能的 DNA 结合蛋白，它的突变可以引起自身免疫病 APECED。AIRE 基因包括两个 PHD 型锌指域和一个 DNA 结合区域 SAND，pHD 型锌指有与大分子（核酸或蛋白质）结合的特性。

#### 6.1.3.2 锌指蛋白与移植免疫

肝脏移植后产生的急性排斥炎症反应和锌指蛋白 A20 有关。锌指蛋白 A20 是一种最早发现于人脐静脉内皮细胞的 TNF-α 诱导性基因产物，其表

达需要 NF-κB 激活，但对 NF-κB 激活起负反馈抑制作用，抑制血管内皮细胞激活。A20 抑制 CD40-CD40L 结合诱发的内皮细胞激活，而 CD40-CD40L 结合是诱发移植排斥的机理之一。A20 还能通过抑制 NF2κB 激活而产生抗炎功效，保护肝细胞并促进肝再生。

#### 6.1.3.3 锌指蛋白与肿瘤免疫

ZNF165 基因是锌指蛋白家族中的一员，在肝癌、胃癌、结肠癌和肺癌中都能检测到 ZNF165 mRNA 的表达，ZNF165 mRNA 以高水平在肝癌组织与睾丸组织中表达，而在配对的癌旁组织与其他正常组织中仅存在很低水平的表达。在肿瘤免疫中，ZNF165 是一种可以诱导出免疫应答的抗原。锌指蛋白 A20 参与机体炎症反应和细胞凋亡的重要调控。研究表明，A20 在包括乳腺癌耐药株、炎性乳腺癌、肾透明细胞癌和结直肠癌细胞系、前列腺癌羟基喜树碱耐药株、鼠纤维肉瘤细胞 SV80 等多种人类肿瘤细胞系和临床标本中过表达。

#### 6.1.3.4 锌指蛋白与白血病

早幼粒细胞白血病锌指基因（promyelocytic leukemia zinc finger，PLZF）与白血病发病密切相关。PLZF 蛋白的 C 端含有 9 个 $C_2H_2$ 型锌指结构，可以和 DNA 结合。转录因子 Sp1 是存在于哺乳动物细胞中的一种基因调控锌指蛋白。Sp1 能调控早幼粒细胞白血病基因（promyelocytic leukemia，PML）的转录，降解 PML-RARα 蛋白，诱导 APL 细胞分化；能调控白血病相关蛋白 EEN（extra eleven nineteen）的表达，限制 EFN 的过量表达而达到治疗白血病的效果；对 K562 白血病细胞增殖、凋亡及耐药性产生影响。

## 6.2 锌指核酸酶

### 6.2.1 核酸酶

在高等动植物中核酸的分解需要核酸酶的参与，核酸酶作用于核苷酸之间的磷酸二酯键。不同来源的核酸酶，相同来源不同种类的核酸酶，其专一性和作用方式都有不同。

根据作用底物的不同，可以分为核糖核酸酶、脱氧核糖核酸酶。只能作用于 RNA 的核酸酶，称为核糖核酸酶（RNase）；只能作用于 DNA 的核酸酶，为脱氧核糖核酸酶（DNase），有些核酸酶既能作用于 RNA 也能作用于 DNA，统称为核酸酶（nuclease）。根据切割位置的不同，又可将核酸酶分为核酸外切酶（exonuclease）和核酸内切酶（endonuclease）。

核酸外切酶（exonuclease）的作用是从分子链的末端顺次水解磷酸二

酯键生成单核苷酸。可大致分为 3′→5′核酸外切酶和 5′→3′核酸外切酶。例如蛇毒磷酸二酯酶是 3′→5′核酸外切酶，嗜酸乳杆菌（*Lactobacillus acidophilus*）核酸酶就是 5′→3′核酸外切酶。

核酸内切酶（endonuclease）是水解核酸内部磷酸二酯键来切断核酸。可分为限制性核酸内切酶和非特异性核酸内切酶。一般都不具碱基特异性，但也有限制性内切酶能够识别并切断特定的碱基或碱基序列的酶。限制性核酸内切酶（restriction endonuclease）简称限制酶，是在细菌细胞内存在的一类能特异性识别并水解外源双链 DNA 的核酸内切酶，可用于特异切割 DNA，常作为工具酶。其能从 DNA 分子中间水解磷酸二酯键，从而切断双链 DNA。它们不同于一般的脱氧核糖核酸酶，切点严格，要求专一的核苷酸顺序。在科研中，难以深入研究的 DNA 大分子，可以用限制性内切酶切割成特定的小片段来进行分析研究。限制性核酸内切酶的发现，为分子生物学的研究开辟了新的途径，已成为必不可少的工具酶。

### 6.2.2 锌指核酸酶

锌指核酸酶由锌指蛋白和一个非特异性核酸内切酶构成。锌指蛋白作为锌指核酸酶的 DNA 识别域，使得锌指核酸酶具有特异性。每个锌指结构识别并结合一个特异的三联体碱基，改变锌指结构的氨基酸序列，可以改变锌指核酸酶的特异性，多个锌指结构可以串联起来，增强了锌指核酸酶的特异性和可塑性。1996 年，Kim 等合成了第一个锌指核酸酶，将锌指蛋白与非特异性核酸酶 Fok Ⅰ连接，组合成锌指核酸酶单体，其二聚体具有活性。其中的非特异性核酸内切酶 Fok Ⅰ是来自海床黄杆菌的一种限制性内切酶，与锌指核酸酶连接的是其 C 端的 96 个氨基酸残基组成的 DNA 剪切域。Fok Ⅰ只在二聚体状态时才有酶切活性，当两个锌指核酸酶的识别位点距离 6～8bp 时，会产生酶切作用。

锌指核酸酶技术应用于实际的科研或医疗，还有许多问题需要解决。首先是锌指核酸酶的特异性。二聚体的形成是独立的，同源二聚体和异源二聚体的形成都可以形成 DNA 的剪切，这种非特异性会造成毒性。Miller 等和 Szczeoek 等在这方面做了很多工作，构建了一系列偏爱异二聚体形成的 Fok Ⅰ酶，但这种情况仍不能完全避免，还需要继续研究。其次，引入的锌指核酸酶蛋白是不是会引起免疫系统的反应尚不明确，到目前这一技术只能取自身细胞，对其进行体外操作，再注回体内。锌指核酸酶的设计相对于其他蛋白的设计更加有利，其特异性较好且结构保守，在治疗人类遗传疾病上有极大的应用价值。经人工设计自然界不存在的蛋白质，用于科学实验、基因治疗和调控，为生物学及医学研究开辟了新的领域。

## 6.3 锌指核酸酶的设计

ZFN 是作为某种人工合成的限制性内切酶对特定的基因起作用。其研究方法主要是通过设计 ZFN 中的 DNA 结合区域（DNA binding domain，DBD）来识别特定的 DNA 序列，通常为特定的目的基因序列，进而与非特异性核酸内切酶融合设计，达到特异性核酸内切酶的作用。如果把锌指核酸酶应用于基因治疗，则主要有以下实验步骤：首先找出疾病基因突变位点，即确定靶位点，然后根据靶位点进行锌指核酸酶的设计并检测锌指核酸酶结合 DNA 的能力。把检测合格的锌指核酸酶导入到细胞内，导入成功的细胞注入生物体内，进行生物实验。

### 6.3.1 靶位点的确定

锌指核酸酶（ZFN）由一个 DNA 识别域和一个非特异性核酸内切酶构成。DNA 识别域是由一系列 Cys2-His2 锌指蛋白（zinc-fingers）串联组成（一般 3～4 个），每个锌指蛋白识别并结合一个特异的三联体碱基。由于锌指核酸酶必须形成二聚体（图 6.2）才具有活性，两个锌指 DNA 结合位点就识别 18 个碱基，加上中间间隔的 6 个碱基一共是 24 个碱基，也就是说，要在基因上寻找一段 24bp 的 DNA 序列作为锌指核酸酶的靶位点。查找基因序列的相关信息，可以在 UCSC（University of California at Santa Cruz，http：//genome. ucsc. edu）及美国国家生物信息中心（http：//www. ncbi. nlm. nih. gov）上查找。早在 2003 年，人类全基因组测序已经完成，可以在 UCSC 的网站上查到。UCSC 主要提供的是基因组的参考序列和相关的分析服务，还提供了 13 个物种中的 CFTR 区域，分为几个不同的部分，包括 Genome Browser、Gene Sorter、Bla 和 Table Browser，分别提供了基因的表达信息、同源物、在染色体上所处的位置，对应于基因组中的序列包括内含子、外显子等信息。

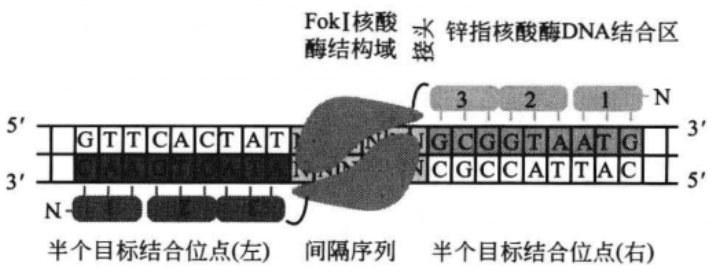



图 6.2 锌指核酸酶二聚体结构

## 6.3.2 ZFN 蛋白序列的设计及优化

### 6.3.2.1 ZFN 序列的获取

登陆 Scripps Research Institute 的 ZF tools 服务器 http：//www. scripps. edu，提交查找好的靶位点，设定相关参数，可获取 ZFN 的序列。

### 6.3.2.2 ZFN 核酸序列优化

使用 Primer Premier5. 0，把 ZFN 氨基酸序列逆向翻译为核酸序列。对所有的密码子进行优化，优化遵循条件：

① 考虑到设计的锌指核酸酶以后要做基因表达（一般用真核表达），要对设计好的核酸序列进行筛选，尽量选择在真核及原核细胞内均可有效表达的密码子，对大肠杆菌内较难表达的密码子的表达阈值设定为不低于 30%，以利于后期原核表达的进行。从密码子偏爱性方面进行分析，所用工具为网站 http：//gcua. schoedl. de/上的 codon vs. usage table。

② 利用密码子的简并性，尽量减少整个序列中局部重复序列的产生，以利于全基因合成及后期基因克隆操作。

### 6.3.2.3 ZFN 序列的特征提取及结构建模

利用 PSIPRED（http：//www. bioinf. cs. ucl. ac. uk/psipred），对 ZFN 蛋白进行特征提取及结构分析。PSIPRED 是基于神经网络算法的蛋白质二级结构预测软件，平均准确率为 78%，它是一种较为准确的蛋白质二级结构预测方法。利用 Swiss-Model（http：//swissmodel. expasy. org/）和 Deep View（http：//www. expasy. org/spdbv）进行 ZFN 的三级结构同源建模。SWISS-MODEL 是一个自动化的蛋白质比较建模服务器，DeepView 是一个整合工具，用于观察和分析蛋白质结构和模型。Swiss-Model 结构合理性评估有三种方式，分别是原子间经验平均势函数 Anolea、Gromos 分子动力学以及表示蛋白质残基与结构环境匹配的 Verify3D 评估。

## 6.3.3 锌指核酸酶设计举例

以镰刀形贫血症为例，设计用于基因治疗的锌指核酸酶。镰刀形红细胞贫血症是一种“分子病”，即分子结构（特别是蛋白质分子结构）发生遗传性变化而造成的病变。

### 6.3.3.1 确定靶位点

异常血红蛋白 β 链的第 6 位谷氨酸被缬氨酸所代替，因此我们先要找出血红蛋白 β 链的基因。登陆 UCSC 网站主页（http：//genome. ucsc. edu）进行检索，如图 6. 3。

点击图 6.3 中的“Genome Browser”按钮，进入图 6.4 所示界面。相应的设置为：Human、July 2003、HBB、620。

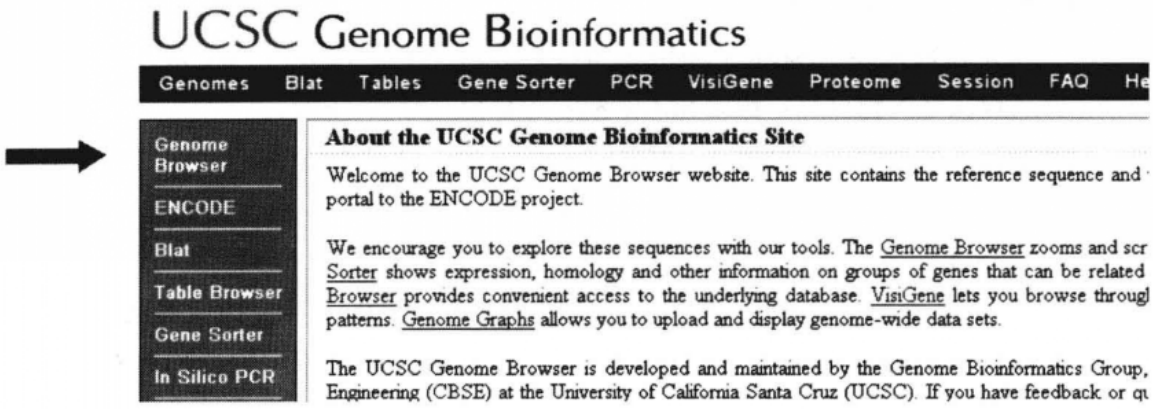



图 6.3　UCSC 数据库网站主页

点击图 6.4 中的“submit”，进入图 6.5 所示界面。图中显示的是人类 11 号染色体的一部分，包含 HBB 基因。

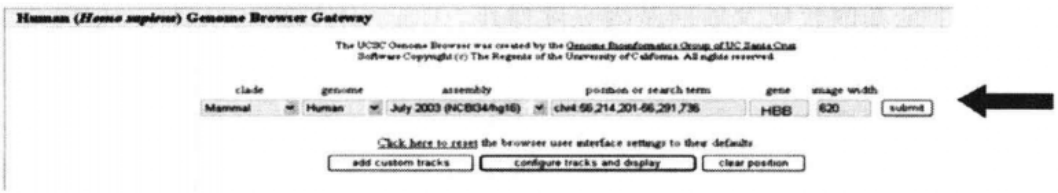



图 6.4　寻找人类血红蛋白 β 链 DNA 的参数设置

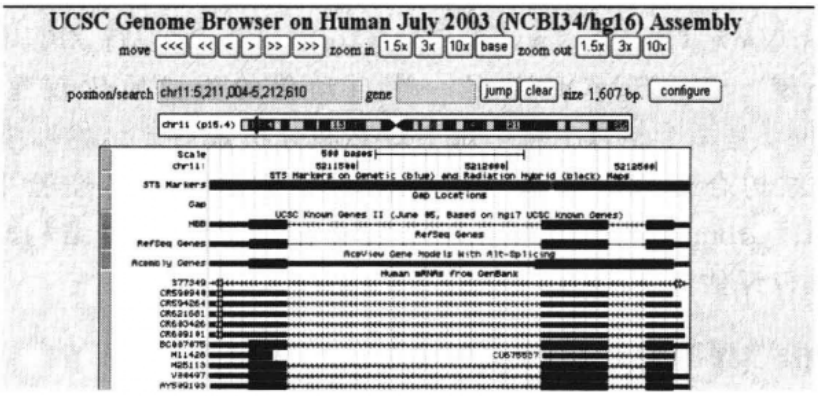



图 6.5　11 号染色体的一部分，包含了人血红蛋白 β 链（HBB）基因

点击图 6.5 中的“HBB”，可以获取 HBB 的详细信息，如图 6.6。点击图 6.6 中的“Genomic Sequence”可以获取 HBB 的 DNA 序列，如图 6.7。

在图 6.5 所示界面的下方找到“Variation and Repeats”，如图 6.8 设定相应的参数。点击图 6.8 中“refresh”按钮，主界面就会出现更多的信息，如图 6.9。

点击图 6.9 中的“rs334”，可进入图 6.10 所示界面，由图 6.10 可以知道突变位点的相关信息。

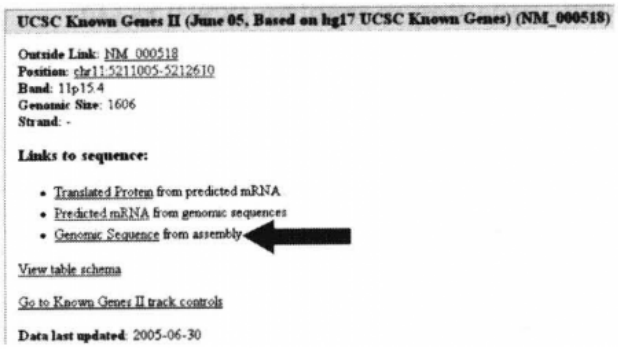



图 6.6　人血红蛋白β链（HBB）基因的详情

图 6.7　人血红蛋白β链（HBB）的基因序列

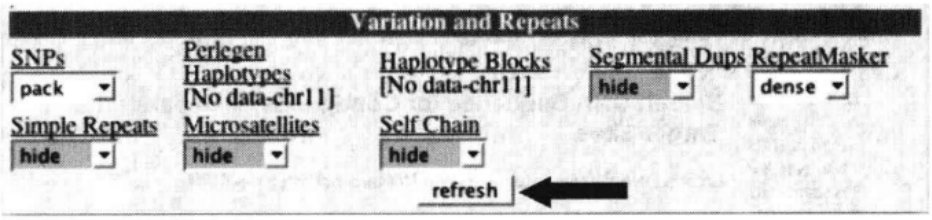



图 6.8　查找变异和重组部分的基因

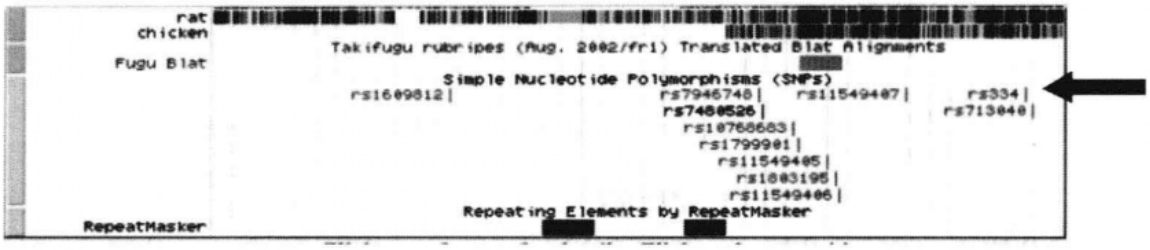



图 6.9　人血红蛋白 HBB 基因的多态性

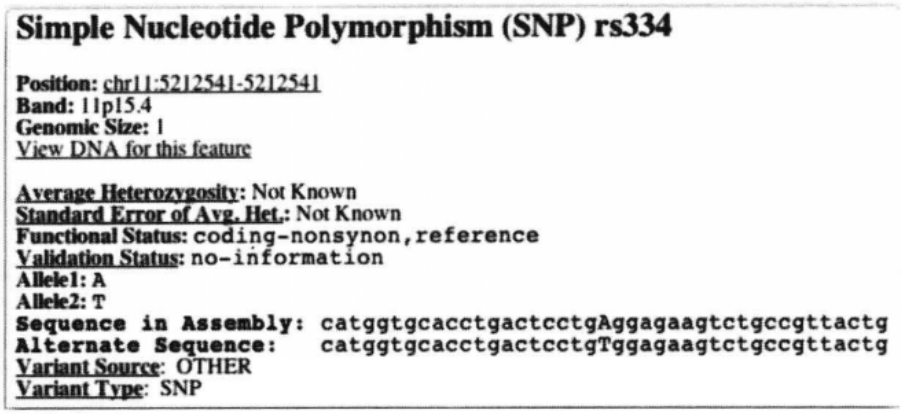



图 6.10 HBB 的突变基因

### 6.3.3.2 ZFN 序列的获取

由于镰刀形细胞贫血病的突变位点存在于血红蛋白 β 链的第一个外显子区域，我们考虑将第一个外显子整个切除，然后导入外源正常基因进行修复。即要设计两个靶位点：一个位于基因上游 100bp 左右；另外一个位于第一个内含子区。这样做的好处在于，即使在 DNA 的接口处引入突变，也不会对细胞产生太大的影响，因为在血红蛋白 β 链基因上游 100bp 处没有其他基因，而内含子最终是会被剪切掉的。在这里，我们只举例设计第一个靶位点的锌指核酸酶。目前，有专门设计锌指蛋白靶位点的网站 http://www.scripps.edu（图 6.11）。在网页中输入相应的 DNA 序列，设定好相应的要求，即可获得一些可以作为靶位点的 DNA 序列。

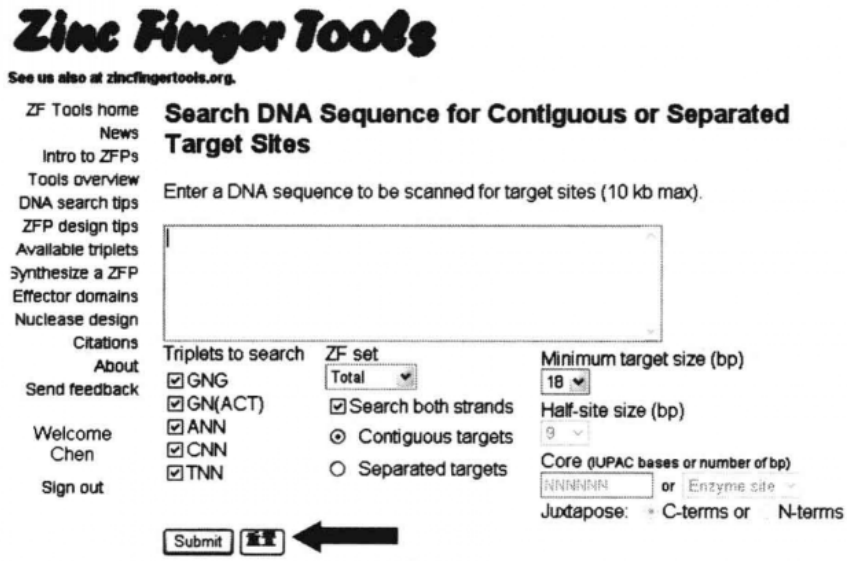



图 6.11 主页和检索工具

这次以血红蛋白β链基因上游100bp内靶位点为例进行设计。在网页中输入血红蛋白β链基因上游100bp的序列：

tgtggagccacaccctagggttggccaatctactcccaggagcagggagggcaggagccagggctgg gcataaaagtcagggcagagccatctattg ctt

选择“Separated targets”，点击“Submit”，得到很多靶位点，如图6.12。其中，红色的碱基（图中标以斜体）是有迭代问题的，不适合作为锌指核酸酶的靶位点，如图6.12中所列的第四个靶位点中的“GAG”。在众多的备选靶位点选取得分高的靶位点，如图6.13所示，选取了得分最高的三个靶位点用于后续实验。其中蓝色标记的碱基（图中标以黑体）为锌指的识别位点，划横线的碱基为核酸酶Fok Ⅰ的作用位点。锌指核酸酶的核酸酶结构域通常选取Fok Ⅰ核酸酶结构域。

**Forward Strand:**
Number of sites found: **19**

⊟ Coverage map. The input sequence with every targetable base colored (**top** and **bottom** strands):

```
5'-TGTGGAGCCA CACCCTAGGG TTGGCCAATC TACTCCCAGG AGCAGGGAGG GCAGGAGCCA GGGCTGGGCA-3' 70
3'-ACACCTCGGT GTGGGATCCC AACCGGTTAG ATGAGGGTCC TCGTCCCTCC CGTCCTCGGT CCCGACCCGT-5'
```

Click **here** for a diagram of separated target sites like those used in nuclease design.

| Spans residues (relative to top strand) | Score (Base) | TSO Issues (Triplets in red) | Parse | ZFP | Half sites and core sequence (Mouse over (or click on) triplet for profile) |
|---|---|---|---|---|---|
| 32 | 60.83 | No | ⇔ | ▶ | 5'-CAC ACC CTA GGGTTG GCC AAT CTA-3' |
| 9 | 35.31 | No | ⇔ | ▶ | 3'-GTG TGG GAT CCCAAC CGG TTA GAT-5' |
| 40 | 45.32 | No | ⇔ | ▶ | 5'-AGG GTT GGC CAATCT ACT CCC AGG-3' |
| 17 | 60.83 | No | ⇔ | ▶ | 3'-TCC CAA CCG GTTAGA TGA GGG TCC-5' |
| 44 | 53.11 | No | ⇔ | ▶ | 5'-TTG GCC AAT CTACTC CCA GGA GCA-3' |
| 21 | 49.47 | No | ⇔ | ▶ | 3'-AAC CGG TTA GATGAG GGT CCT CGT-5' |
| 45 | 61.11 | Yes | ⇔ | ▶ | 5'-TGG CCA ATC TACTCC CAG *GAC* CAG-3' |
| 22 | 58.39 | No | ⇔ | ▶ | 3'-ACC GGT TAG ATGAGG GTC CTC GTC-5' |

图6.12　上游靶位点检索的部分结果

| | | | | | |
|---|---|---|---|---|---|
| 58 | 57.67 | No | ⇔ | ▶ | 5'-CCC AGG AGC AGGGAG GGC AGG AGC-3' |
| 35 | 58.56 | No | ⇔ | ▶ | 3'-GGG TCC TCG TCCCTC CCG TCC TCG-5' |
| 61 | 57.67 | No | ⇔ | ▶ | 5'-AGG AGC AGG GAGGGC AGG AGC CAG-3' |
| 38 | 58.56 | No | ⇔ | ▶ | 3'-TCC TCG TCC CTCCCG TCC TCG GTC-5' |
| 80 | 56.59 | No | ⇔ | ▶ | 5'-GCC AGG GCT GGGCAT AAA AGT CAG-3' |
| 57 | 60.24 | No | ⇔ | ▶ | 3'-CGG TCC CGA CCCGTA TTT TCA GTC-5' |

图6.13　上游靶位点检索得分最高的三组结果

点击箭头标志，可以进入关于对应该位点的锌指域的氨基酸序列的网页，如图6.14。由于我们在基因上游100bp选取了三组，最终我们得到总共6条氨基酸序列（表6.1）。其中粗体标记的氨基酸序列对应锌指。

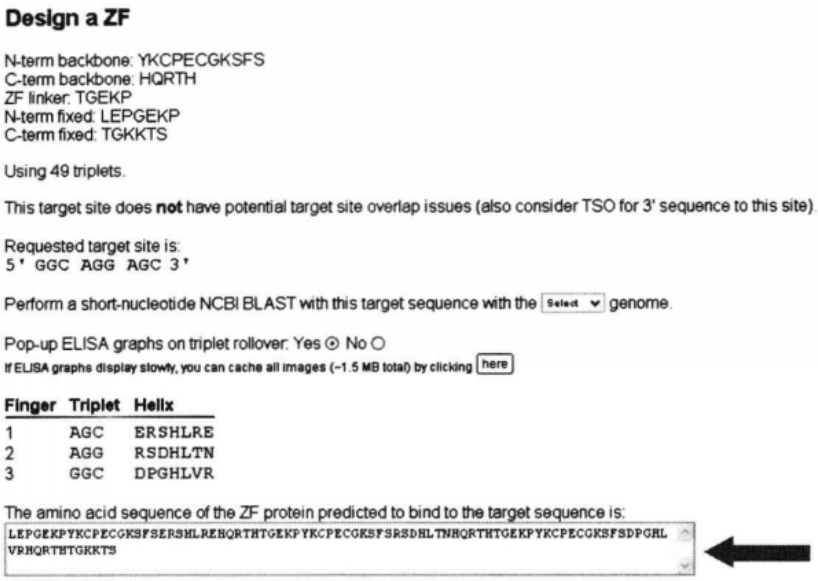



图 6.14　含三个锌指结构域的氨基酸序列

**表 6.1　上游三组锌指域的氨基酸序列**

| 跨度 | 预测的锌指蛋白的氨基酸序列 |
|---|---|
| 58 | LEPGEKPYKCPECGKSFS**ERSHLRE**HQRTHTGEKPYKCPECGKSFS**RSDHLTNH**QRTHTGEKPYKCPECGKSFS**DPGHLVR**HQRTHTGKKTS |
| 35 | LEPGEKPYKCPECGKSFS**RSDKLVR**HQRTHTGEKPYKCPECGKSFS**TKNSLTE**HQRTHTGEKPYKCPECGKSFS**TSGELVR**HQRTHTGKKTS |
| 61 | LEPGEKPYKCPECGKSFS**RADNLTE**HQRTHTGEKPYKCPECGKSFS**ERSHLRE**HQRTHTGEKPYKCPECGKSFS**RSDHLTN**HQRTHTGKKTS |
| 38 | LEPGEKPYKCPECGKSFS**TKNSLTE**HQRTHTGEKPYKCPECGKSFS**TSGELVR**HQRTHTGEKPYKCPECGKSFS**TKNSLTE**HQRTHTGKKTS |
| 80 | LEPGEKPYKCPECGKSFS**RADNLTE**HQRTHTGEKPYKCPECGKSFS**HRTTLTN**HQRTHTGEKPYKCPECGKSFS**QRANLRA**HQRTHTGKKTS |
| 57 | LEPGEKPYKCPECGKSFS**DPGHLVR**HQRTHTGEKPYKCPECGKSFS**TKNSLTE**HQRTHTGEKPYKCPECGKSFS**ERSHLRE**HQRTHTGKKTS |

#### 6.3.3.3　ZFN 核酸序列优化

把 ZFN 的氨基酸序列转化成核苷酸序列。用 Primer Premier5.0，把 ZFP 氨基酸序列逆向翻译为核酸序列。打开 Primer Premier5.0 软件，在 File 中选择 New 选项中的 Protein Sequence 选项，会弹出对话框（图 6.15）。在空白处输入蛋白质的氨基酸序列（以第一条氨基酸序列为例），再点击“DNA”键，就可以得到其对应的核苷酸序列（图 6.16）。其中 G、C 分别代表鸟嘌呤和胞嘧啶，A、U 分别代表腺嘌呤和尿嘧啶。N 可以是 G、C、A、U 中任何一种碱基；Y=C 或 U；R=G 或 A。

去掉稀有密码子。因为有不确定的碱基：N 可以是 G、C、A、U 中任何一种碱基，Y=C 或 U，R=G 或 A。根据表达载体并尽量避开稀有密码子。

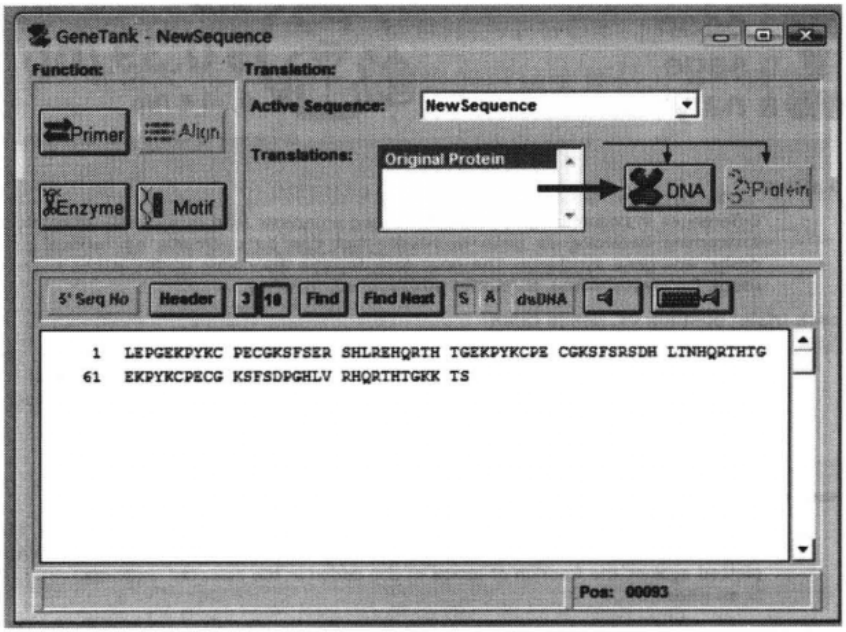



图 6.15　在 Primer Premier5.0 中输入蛋白质序列

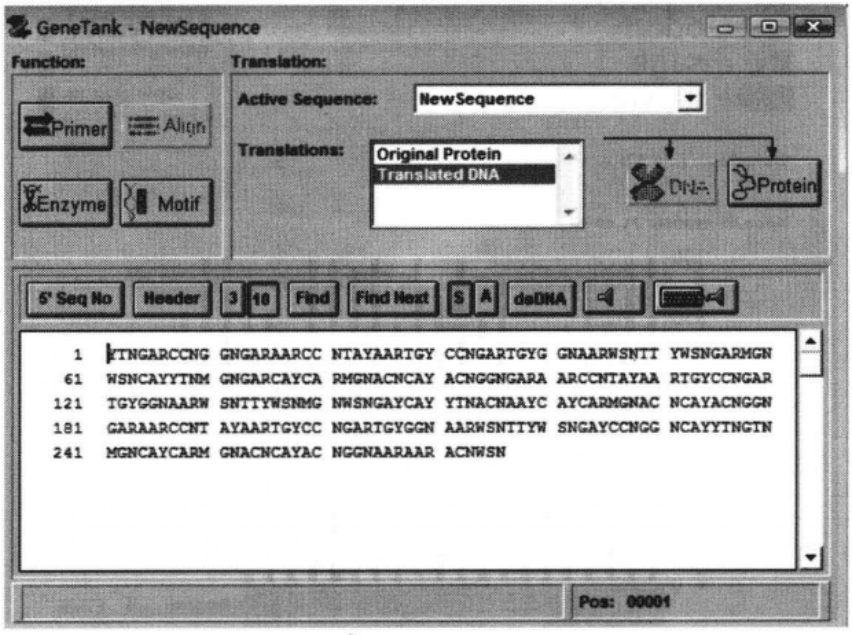



图 6.16　从 Primer Premier5.0 中得到的核苷酸序列

登陆网站 http://gcua.schoedl.de/（图 6.17），点击 codon vs. usage table。进入页面后，输入相应的密码子（以 AAA 为例），选择表达载体（以大肠杆菌为例），设置完成后点击“Submit”得到密码子的使用频率（图 6.18）。红色的为我们输入的密码子的使用频率。按照这种方法，对编码 ZFN 的密码子进行确定和检查。

**purpose**

Differences in codon usage preference among organisms lead to a variety of problems concerning heterologous gene expression but can be overcome by rational gene design and gene synthesis. The gcua tool displays the codon quality either in codon usage frequency values or relative adaptiveness values.

**each triplet position vs. usage table**

Submit a DNA sequence in raw or fasta format and choose a codon usage table. The sequence will be splitted in codons and the fraction of usage of each codon in the selected organism will be represented as one column. Additonal to the listed codon usage tables, you can submit your own by pasting in a http address.
**an example...**

**each codon vs. usage table**

Submit a DNA sequence in raw or fasta format and choose a codon usage table. The fraction of usage of each codon in the submitted sequence will be computed and plotted against the fraction of usage of the codon in the selected organism.
**an example...**

图 6.17　密码子数据库及密码子偏好性分析软件主页

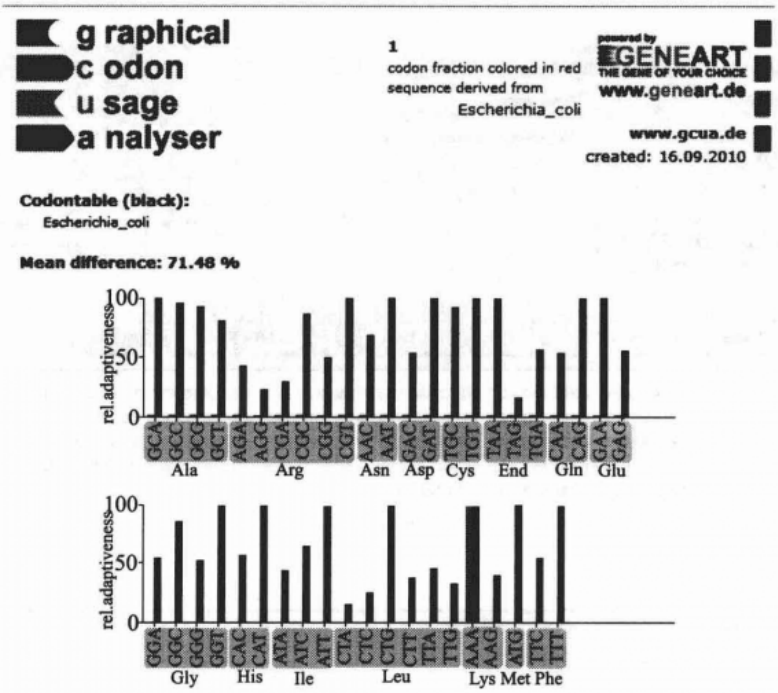



图 6.18　各种密码子在大肠杆菌中的使用频率（见彩图）

### 6.3.3.4　ZFN 序列的特征提取及结构建模

构建的锌指核酸酶二级结构的模拟及分析。登陆 PSIPRED 的页面 http：//bioinf. cs. ucl. ac. uk/psipred/，如图 6.19。在页面中填入 ZFN 蛋白质的序列，自己的电子邮箱，点击页面中间的“Predict”命令，就可以完成预测。以第 1 条 ZFN 氨基酸序列为例，输入氨基酸序列：LEPGEK-

PYKCPECGKSFSERSHLREHQRTHTGEKPYKCPECGKSFSRSDHLTNH
QRTHTGEKPYKCPECGKSFSDPGHLVRHQRTHTGKKTS，就可以得到预测结果，如图 6.20。结果包括预测的可信度评估及各个残基的构象。

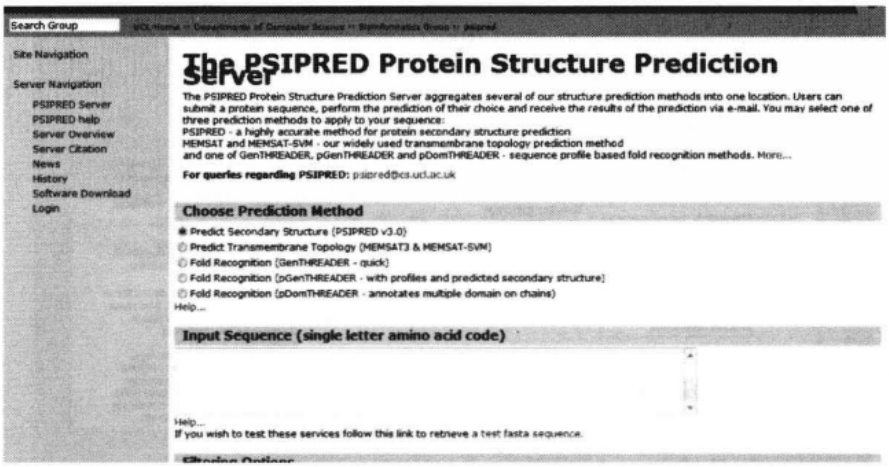



图 6.19　PSIPRED 蛋白质二级结构预测主页

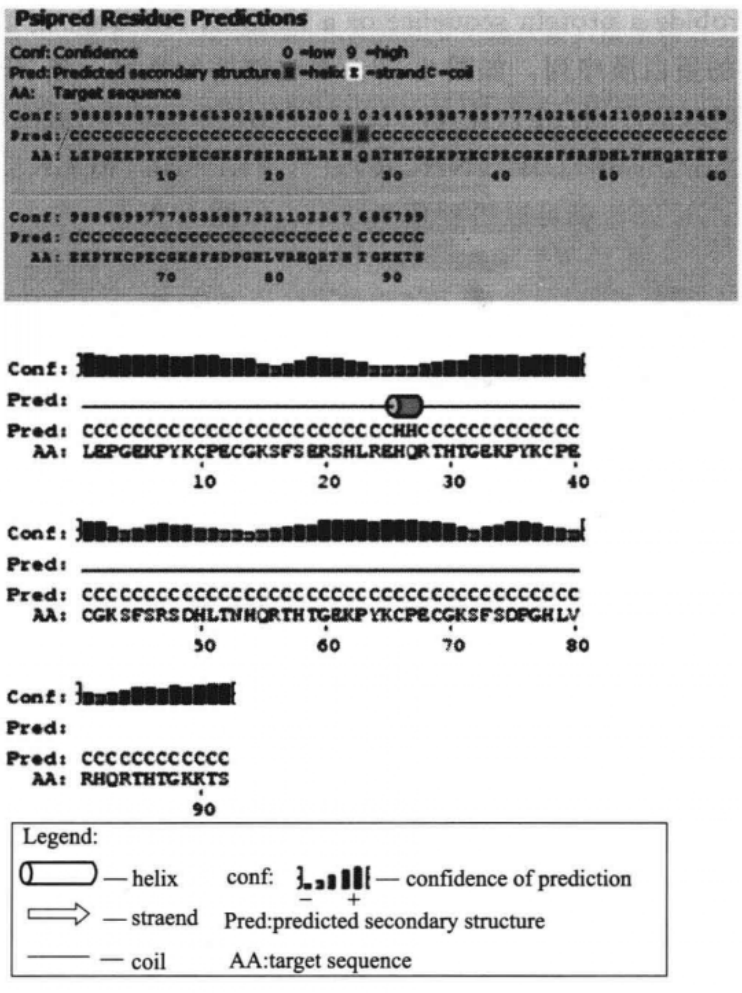



图 6.20　PSIPRED 蛋白质二级结构预测结果

构建的锌指核酸酶三级结构的模拟及分析。登陆 SWISS-MODEL 服务器的 web 界面，如图 6.21。其共有三种模式：自动模式（automated mode）、联配模式（alignment）、项目模式（project mode）。本次选取自动模式。

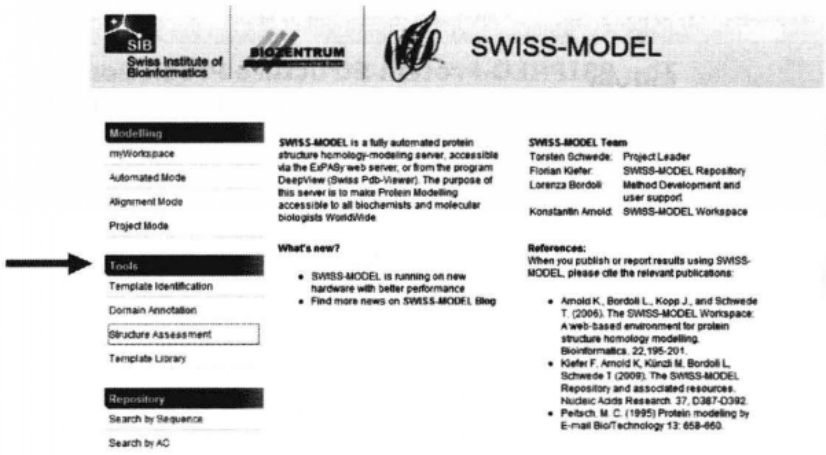



图 6.21　SWISS-MODEL 服务器页面

在 Probide a protein sequence or a UniProt AC Code 对话框中输入需要预测结构的蛋白质序列，如图 6.22。本次使用的是第一条 ZFN 的氨基酸序列，输入氨基酸序列：LEPGEKPYKCPECGKSFSERSHLREHQRTHTGEKPYKCPECGKSFSRSDHLTNHQRTHTGEKPYKCPECGKSFSDPGHLVRHQRTHTGKKTS，就可以得到预测结果，如图 6.23。

图 6.22　SWISS-MODEL 服务器工作界面

在结果中，建模范围为 4.58，目标序列与模板序列的相似性为 87.273%。蛋白质的结构模型也可以通过 Deep View 查看，点击“download”，下载 pdb 格式的文件，然后下载 Deep View 软件，打开。

在出现建模结果的同时，还有目标蛋白与模板结构的序列联配信息，如图 6.24；结构的合理性评估信息，如图 6.25。对建模结果的评估共有三种，用 on 和 off 来控制，本次实验选取了全部三种评估。

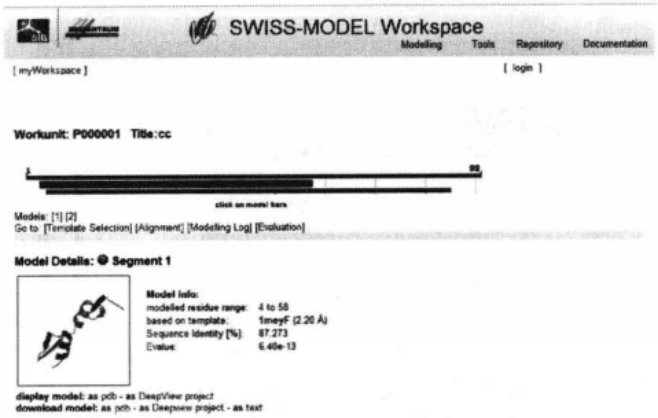



图 6.23　SWISS-MODEL 服务器的建模结果

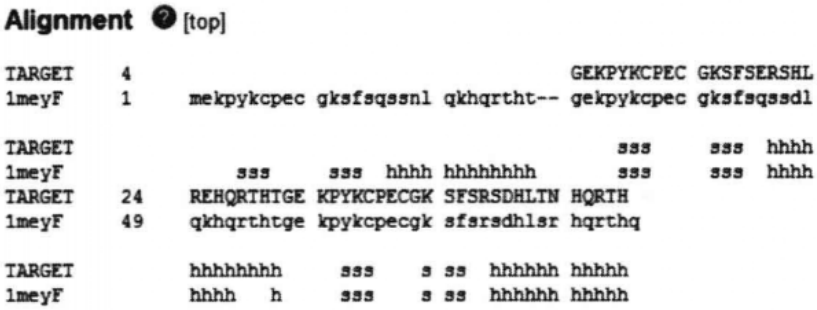



图 6.24　SWISS-MODEL 服务器建模过程中，目标序列与模板序列的联配

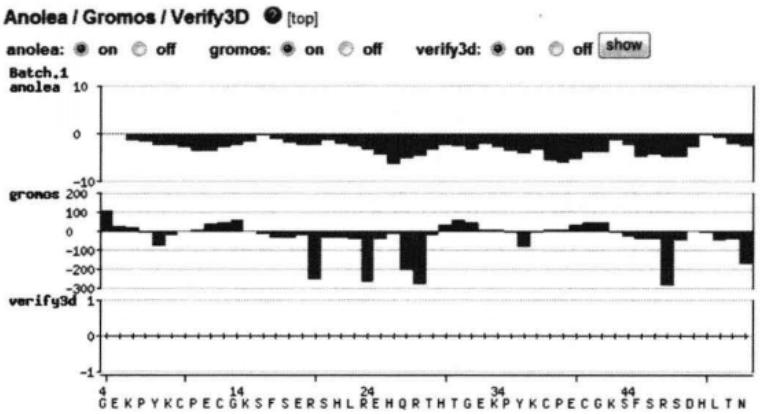



图 6.25　SWISS-MODEL 服务器对建模结果的部分评估

下载的 pdb 格式的文件可以用 Deep View 软件打开，如图 6.26，更方便对结果的查看和分析。本次试验用 Deep View 软件打开的就是第 1 条 ZFN 氨基酸序列。图 6.26 中，红色区域为 α 螺旋，蓝色为 β 折叠，绿色为无规则卷曲。因为 SWISS-MODEL 服务器建模范围为 4.58，并没有对 92 个氨基酸全部建模，这在 Deep View 软件中可以清楚地看到。

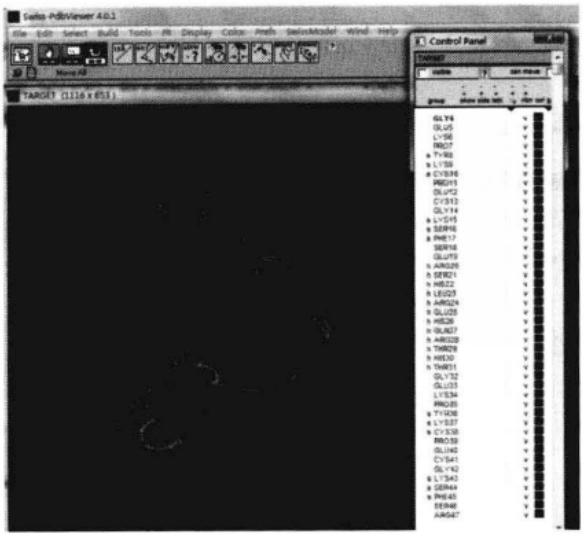


图 6.26　Deep View 软件工作界面（见彩图）

#### 6.3.3.5　数据分析

对设计的不同组的锌指核酸酶进行二级结构和三级结构模拟的结果进行评估，确定合适的锌指核酸酶用于后续试验。可选择在结果中包含一份蛋白质校验工具 WhatCheck（http：//www. cmbi. kun. nl/gv/whatcheck ）的报告。另外，也可以利用服务器上的 ANOLEA 原子平均势能的 web 界（http：//swissmodel. expasy. org/anolea）检查模型的质量。

随着科技的发展，模拟天然蛋白质的结构，人工设计自然界不存在的蛋白质对特定基因进行调控，为现代基因治疗的发展提供了新的思路。锌指蛋白依靠其特殊的锌指结构域来识别核酸序列，在生物体中起到了重要作用。虽然目前锌指核酸酶技术还处在进一步完善阶段，但在转基因动植物的研究和应用方面拥有巨大的潜力。利用锌指核酸酶对生物基因组进行靶向修饰，能够缩短动植物优良新品种的培育时间。更重要的是，锌指核酸酶在治疗人类遗传疾病方面具有极大的应用价值。利用锌指核酸酶对人体细胞进行靶向修饰，可纠正某些遗传缺陷，为克服许多现阶段无法治愈的疾病提供一个强大的治疗工具。

## 参考文献


[1]　Miller J，McLachlan A D and Klug A. The EMBO Journal，1985，4：1609-1614.

[2]　Geurts A M，et al. Science，2009，325：433.

[3]　Gommans W M，Haisma H J and Rots M G. Journal of Molecular Biology，2005，354：507-519.

[4]　魏勇等．生物医学工程学杂志，2008，25：662-667.

[5]　赵楠等．生物技术通讯，2009，20：131-134.

[6]　李荣平等．免疫学杂志，2009，25：357-360.

[7]　李晓波，张俊武．中国生物化学与分子生物学报，2009，25 ：206-211.

[8] Mandell J G and Barbas Ⅲ C F. Nucleic Acids Research, 2006, 34: 516-523.
[9] Arnold K, et al. Bioinformatics, 2006, 22: 195-201.
[10] Schwede T, et al. Nucleic Acids Research, 2003, 31: 3381-3385.
[11] Guex N and Peitsch M C. Electrophoresis, 1997, 18: 2714-2723.
[12] Shukla V K, et al. Nature, 2009, 459: 437-441.
[13] Cathomen T and Joung J K. The American Society of Gene Therapy, 2008, 16: 1200-1207.
[14] 岳俊杰等. 蛋白质结构预测实验指南. 化学工业出版社, 2010.
[15] 加州大学圣克鲁兹分校（UCSC）: http: //genome. ucsc. edu
[16] 锌指蛋白设计工具: http : //www. scripps. edu
[17] 查找稀有密码子的网站: http: //gcua. schoedl. de/
[18] 蛋白质二级结构预测工具: http: //bioinf. cs. ucl. ac. uk/psipred/
[19] 蛋白质三级结构数据库: http: //www. rcsb. org/pdb/
[20] 蛋白质比较建模服务器 Swiss-Model: http : / /swissmodel. expasy. org/

# 第7章 诱导多能干细胞的关键转录因子 Oct4 结构模拟与设计

仲从浩　李荣秀

## 7.1 导读

近年来，诱导多能干细胞研究热潮持续高涨。本章将以诱导多能干细胞的关键转录因子 Oct4 的结构模拟为例，说明如何使用在线服务器 SWISS-MODEL 对蛋白质进行同源建模。

干细胞是一种具有自我更新能力和分化潜能的细胞，可根据分化潜能的大小不同或者来源不同进行分类。其中，胚胎干细胞由于在人类组织或器官修复及替代等医学问题上具有巨大的临床应用前景而备受关注。但胚胎干细胞临床应用研究在一定程度上受限于伦理问题及免疫排斥问题。

近年来，诱导多能干细胞研究热潮持续高涨，该技术的出现有效地回避了胚胎干细胞研究所面对的上述两大难题，推动干细胞临床应用向前迈进一大步。在诱导多能干细胞过程中，转录因子 Oct4 发挥着关键作用。

蛋白质的结构决定功能。对蛋白质结构的解析，无疑能够更深入地了解该蛋白的性质及功能，进而有助于阐明其作用的分子机理。目前对于蛋白质三维结构的研究方法主要可分为两大类：实验方法及计算机预测方法。在计算机预测方法中，同源模拟法是目前为止预测最为成功、准确度最高的方法。SWISS-MODEL 是目前使用最广泛的免费在线同源建模服务器之一。在本章 7.8 节即利用 SWISS-MODEL 中的全自动模式对 Oct4 的 $POU_S$ 结构域进行同源建模。

## 7.2 干细胞

“干细胞”一词最早出现于 19 世纪，1896 年，Wilson 第一次用到“干细

胞”这个名词，专指寄生性蠕虫（如蛔虫生殖系统）的早期细胞。当时人们对干细胞的理解仅限于能够产生子代细胞的一种较原始的细胞。目前认为干细胞是一种具有自我更新（self-renewal）能力和分化（differentiation）潜能的细胞。

根据分化潜能的大小，干细胞可分为全能干细胞（totipotent stem cell）、多功能干细胞（pluripotent stem cell）、多潜能干细胞（multipotent stem cell）和单能干细胞（unipotent stem cell）。根据来源不同，干细胞又可分为胚胎干细胞（embryonic stem cell）和成体干细胞（adult stem cell）。

卵子和精子融合产生受精卵，受精卵在形成胚胎过程中四细胞期之前任一细胞皆是全能干细胞，具有发育成独立个体的能力（图 7.1）。多功能干细胞是全能干细胞的后裔，无法发育成一个个体，但具有发育成多种组织的能力。多潜能干细胞只能分化成特定组织或器官等特定族群（例如血细胞，包括红血细胞、白血细胞和血小板）。单能干细胞只能产生一种细胞类型，但是自更新属性将其与非干细胞区分开。

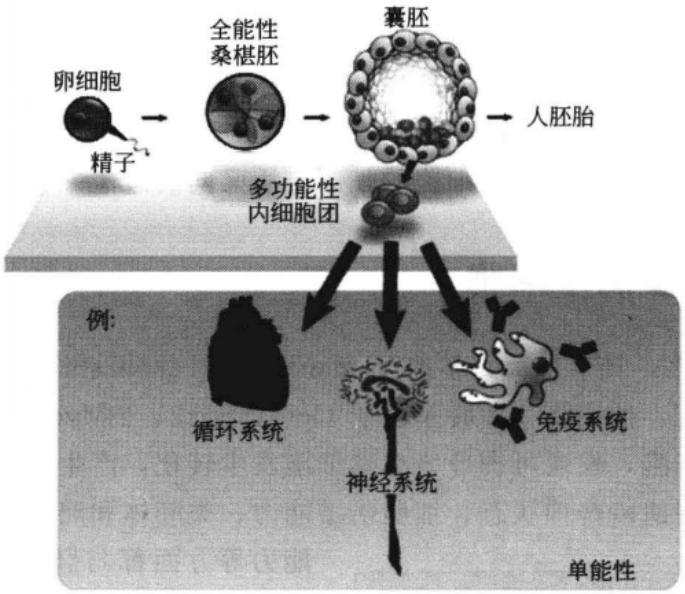



图 7.1　具有不同分化潜能的人类干细胞

胚胎干细胞是在胚胎发育早期的囊胚中，可发育为各种功能细胞的细胞，是所有细胞最初期的形态。成体干细胞亦称成人干细胞，它们存在于成体特定的组织中，可以分化成具特定功能的细胞，例如骨髓干细胞、造血干细胞、神经干细胞。现在医学上常用的成体干细胞大致包括骨髓干细胞、脐带血干细胞及周边血干细胞。

在各种干细胞的研究与应用中，胚胎干细胞最引人注目。最新的研究表明，利用胚胎干细胞自我更新、高度增殖和多向分化的能力，结合现代生物

医学及工程技术，有可能实现人类组织或器官的修复及替代。该领域的研究几乎涉及人体所有的重要组织和器官，也涉及人类面临的大多数医学难题，如：成分输血，意外损伤患者的植皮、关节置换，糖尿病患者的胰腺植入，心脏病患者的瓣膜置换，癌症患者手术切除后组织或器官的替代，大剂量放化疗后造血与免疫功能重建，肝、肾等重要器官损伤或功能衰竭后的置换，部分遗传缺陷性疾病的治疗等。

可胚胎干细胞研究也一直是一个颇具争议的领域。支持者认为这项研究有助于根治很多疑难杂症，是一种挽救生命的慈善行为，是科学进步的表现。而反对者则认为，进行胚胎干细胞研究就必须破坏胚胎（获取胚胎干细胞需要在胚泡阶段破坏胚胎，以提取内细胞团），而胚胎是人尚未成形时在子宫内的生命形式。因此，如果支持进行胚胎干细胞研究就等于是怂恿他人“扼杀生命”，这种行为是不道德的，是违背伦理的。

此外，胚胎干细胞的临床应用本身也存在严重的问题。从胚胎干细胞分化得到的各种细胞和组织如果不是由受体提供的体细胞核移植所得，那么将会与受体发生免疫排斥反应。胚胎干细胞的抗原性可能没有成熟细胞强，但随着胚胎干细胞向成熟细胞的分化，抗原性将逐渐表现出来。

近年来一项蓬勃发展的技术——诱导多能干细胞技术，可以说有效地回避了胚胎干细胞研究所面临的上述两大问题，该技术的出现使干细胞研究登上一个新的平台，并使干细胞的临床应用又向前迈了一大步。

## 7.3 诱导多能干细胞

2006 年，日本京都大学的 Yamanaka 和 Takahashi 在《细胞》杂志上发表一项研究。他们将 4 种转录因子 Oct4、Sox2、c-Myc 和 Klf4 导入小鼠皮肤纤维母细胞，发现可诱导该成熟细胞发生转化，产生的细胞在形态、基因表达、表观遗传修饰状态、细胞倍增能力、类胚体和畸形瘤生成能力及分化能力等方面都与胚胎干细胞极为相似。他们将这种细胞命名为诱导多能干细胞（induced pluripotent stem cells，iPS 细胞）（图 7.2）。

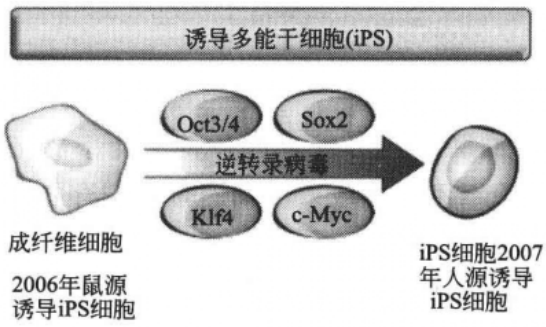



图 7.2　诱导多能干细胞

随后，美国几家实验室也陆续建立 iPS 细胞。2007 年 11 月下旬，美国 Thompson 实验室和 Yamanaka 实验室几乎同时报道，利用 iPS 技术同样可以诱导人皮肤纤维母细胞转变成为几乎与胚胎干细胞完全一样的多能干细胞（图 7.3），所不同的是，Yamanaka 实

验室依然采用了用逆转录病毒导入 Oct4、Sox2、c-Myc 和 Klf4 这 4 种转录因子，而 Thompson 实验室则采用了以慢病毒载体导入 Oct4、Sox2、Nanog 和 Lin28 这 4 种转录因子。

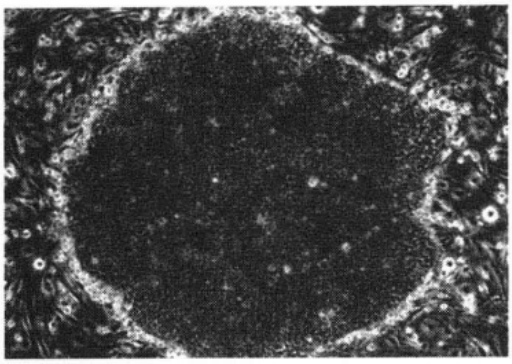


图 7.3　人类 iPS 细胞群落

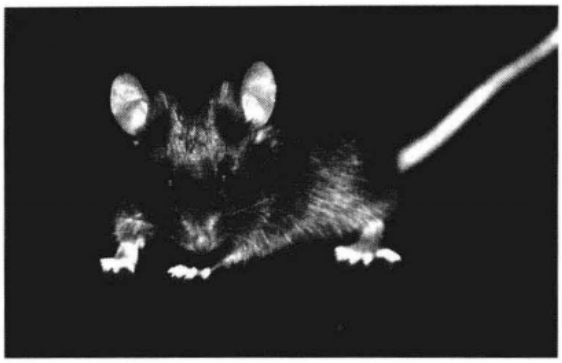


图 7.4　iPS 细胞发育成的小鼠“小小”

2008 年，iPS 研究热潮持续高涨，并取得了多项令人瞩目的进展。哈佛大学 George Daley 实验室利用 iPS 技术把采自 10 名不同遗传病患者身上的皮肤细胞转变为 iPS 细胞，这些细胞将会在建立疾病模型、药物筛选等方面发挥重要作用。哈佛大学另一家实验室则发现利用病毒将 3 种在细胞发育过程中起重要作用的转录因子引入小鼠胰腺外分泌细胞，可以直接诱导其转变成与干细胞极为相似的细胞，并且可以分泌胰岛素。这表明利用诱导重新编程技术可以直接获得某一特定类型的组织细胞，而不必先经过诱导多功能干细胞这一步。

2008 年底，Huangfu 等利用小分子丙戊酸（valproic acid，VPA）和两种转录因子 Oct4 及 Sox2 成功诱导出 iPS 细胞，且诱导效率大大提高。该研究去除了 c-Myc 和 Klf4 这两个有致瘤风险的因子，大大提高了 iPS 细胞应用于临床的可能性。另外，VPA 的引入提示，或许可以选择出合适的小分子物质来完全代替外源基因的导入，从而避免对人类基因组的扰乱，使 iPS 细胞应用于临床变得更加安全可行。

2009 年初，在转录因子的研究方面又取得了新的突破。Kim 等研究表明，只需要 Oct4 这一个转录因子就可以使成年小鼠的神经干细胞重新编程转变为 iPS 细胞。该研究小组将 Oct4 以逆转录病毒为载体转染入小鼠神经干细胞，成功获得 iPS 细胞，这些细胞在体外可以诱导分化为神经干细胞、心肌细胞和生殖细胞。更少因子的使用意味着更少的遗传操作，这将会相应提高 iPS 细胞临床应用的安全性。同年，该小组用功能性的蛋白质（Oct4、Sox2、c-Myc 和 Klf4）取代过表达的四个转录因子，也成功诱导出 iPS 细胞。因为在诱导 iPS 后蛋白质会慢慢降解，所以大大降低了致癌可能性，从而进一步促进该技术的临床应用。

目前猴、大鼠和猪的 iPS 细胞也都成功建立，证明了不同物种之间有着高度的同源性及保守的维持发育多能性的机制。

在这项 iPS 研究热潮中中国后来居上甚至走到了世界的前列。2007 年，中国科学院广州生物医药与健康研究院华南干细胞与再生医学研究所采用逆转录病毒载体导入 4 种转录因子——Oct4、Sox2、c-Myc 和 Klf4，将野生型小鼠成纤维细胞改造成 iPS 细胞。2009 年 1 月，肖磊研究组成功诱导出大鼠 iPS 细胞。同年，中国科学院动物研究所周琪研究组和上海交通大学医学院曾凡一研究组利用 iPS 细胞，通过四倍体囊胚注射得到存活并具有繁殖能力的小鼠“小小”（图 7.4），在世界上首次证明了 iPS 细胞的全能性。“小小”的出生立即入选美国《时代周刊》公布的当年十大医学突破。英国《自然》杂志报道称，中国科学家“为克隆成年哺乳动物开辟了一条全新道路”，并指出“该方法比传统克隆方法更高效、更安全，会引起人们对治疗性克隆的兴趣”。

iPS 技术的出现为研究细胞核重新编程开辟了一条新途径，同时也给再生医学带来了新的希望。iPS 技术证实了既往所认为的处于终末分化的成熟细胞可通过导入外源转录因子基因而发生重新编程，恢复多分化潜能。而且与经典的胚胎干细胞技术和体细胞核移植技术相比，iPS 技术不使用胚胎细胞或卵细胞，因此回避了一直以来争论不休的伦理学问题。此外，iPS 技术可以利用病人自己的体细胞制备专用的干细胞，所以不具有器官移植中的免疫排斥问题。鉴于此，该技术自 2006 年一经报道，便立刻引起科学界甚至媒体界的轰动。2007 年，iPS 技术更是入选美国《科学》杂志年度十大科技突破。

## 7.4 转录因子 Oct4

转录因子 Oct4 发现于 1989 年，这是一种定位于细胞核内，调控哺乳类胚胎发育的关键分子，也是维持胚胎干细胞不分化及自我更新的关键分子。

Oct4 由定位于人类染色体 6p21.3 上的 Pou5f1 基因编码产生。该蛋白是含 POU（Pit-Oct-Unc）结构域的转录因子家族中的一员，属于 POU 蛋白家族 7 个亚族的第Ⅴ亚族，在不同物种间具有高度保守性，含有家族的保守区 POU 结构域。

POU 结构域是 DNA 双向结合结构域，包括两部分：POU 特异域（POU-specific domain，$POU_S$ 结构域）和 POU 同源域（POU-homeo domain，$POU_{HD}$结构域）。$POU_S$结构域位于 N 端，由富含脯氨酸和酸性残基的 75 个氨基酸组成。$POU_{HD}$结构域位于 C 端，由富含脯氨酸、丝氨酸和苏氨酸的约 60 个氨基酸组成。这两个亚区间由长度可变的柔性连接肽（linker）连接，使它们可以通过螺旋-转角-螺旋（HTH）模序结合到 DNA 结

合位点上，激活启动子或增强子区域内带有顺式反应元件的基因表达。后者的特征性结构为 ATGCAAAT 八聚体结构，又称 Oct 结构。Oct4 通过结合含 ATGCAAAT 的八聚体结构而活化相应靶基因，激活或抑制干细胞分化过程中相关基因的表达。POU 结构域以外的氨基酸序列对 Oct4 结合到 DNA 上不是必需的，并且表现出较少的序列保守性。

Oct4 是细胞全能性的标志，也是维持细胞多能性的重要转录因子，通过调控下游靶基因而参与正常发育过程，尤其在早期胚胎发育中起重要作用。它诱导表达的靶基因产物是成纤维母细胞生长因子-4（fibroblast growth factor-4，FGF-4）等，能够通过生长因子的旁分泌作用调节干细胞以及周围滋养层进一步分化。敲除小鼠的 Oct4 基因后，细胞不再分化成内细胞团，而只能发育到囊胚期，小鼠胚胎附植后不久死亡。但在胚泡形成之前抑制 Oct4 功能并不影响卵裂球细胞全能性，这表明早期鼠胚中 Oct4 只有在内细胞团变成滋养层细胞阶段为维持其全能性表型而发挥作用。Oct4 的精确表达是维持胚胎干细胞多能性的重要指标，过表达或低表达都会改变胚胎干细胞的自我更新状态。Oct4 通过与其他转录因子如 Sox2、FoxD3 结合，共同作用于许多下游靶基因的启动子或增强子，正/负调控下游靶基因的表达，从而维持胚胎干细胞的自我更新状态。

近几年，许多研究还发现 Oct4 也表达于多种成体干细胞中，如来源于乳腺、胰腺、肝和肾的干细胞以及毛囊干细胞、间充质干细胞和神经祖细胞等。Hochedlinger 等研究发现异位表达 Oct4 基因可引起上皮细胞异常增生，可能与肿瘤形成有密切的关系。而近几年研究也发现 Oct4 基因表达于膀胱癌、乳腺癌、胰腺癌、骨和软骨瘤中，但在正常组织中不表达。作为胚胎干细胞的特异性基因，Oct4 在肿瘤中的表达提示细胞癌变过程中可能存在胚胎基因的激活，也有研究者认为这是干细胞致癌的一种证据，换句话说，Oct4 基因可能是一个潜在的癌干细胞标记物。

然而，Leendert 等利用 100 多种肿瘤组织芯片对 Oct4 基因进行表达情况筛查时却发现，该基因在非生殖细胞肿瘤中几乎不表达。此外，部分研究者质疑 Oct4 蛋白在非胚胎性肿瘤细胞系和组织中的表达可能是假基因和亚型（isoforms）的影响，甚至认为是实验过程中对照设置不合理，过度曝光造成的。因此，Oct4 基因在成体肿瘤细胞中是否表达尚有争议，这方面的探究工作正在积极开展中。

## 7.5 转录因子 Oct4 结构解析的意义

由于 iPS 技术回避了胚胎干细胞研究所带来的伦理问题，而且其能产生个体特异的细胞群，从而可以避免免疫排斥反应，因而该技术在提出后受到

科学家及临床研究工作者的广泛关注。但是在 iPS 细胞正式进入临床使用之前还有许多疑惑需要解答，而 iPS 细胞重新编程的分子机制便是首先需要弄清楚的。

目前科学家们正从细胞信号通路、small RNA、转录因子及表观遗传等不同角度研究 iPS 细胞重新编程的机理，以及 iPS 细胞自我更新、增殖及分化的机理。在众多研究当中，Oct4 一直被认为是不可或缺的转录因子，在重编程起始过程中起到关键作用。如 Kim 等研究表明：只需要 Oct4 这一个转录因子就可以诱导成年小鼠的神经干细胞重新编程转变为 iPS 细胞。

此外，Oct4 的精确表达也是维持胚胎干细胞多能性的重要指标，过表达或低表达都会改变胚胎干细胞的自我更新状态。科学家们也正积极探究 Oct4 转录因子与相关辅因子协同调控下游基因表达的分子机制。而 Oct4 能否作为肿瘤干细胞的标志，近年来也在热烈探讨中。

蛋白质的结构决定功能。对 Oct4 结构的解析，无疑可以更深入地了解该蛋白质的性质及功能，进而有助于阐明其与辅因子及下游靶基因作用的分子机理。这对揭示 Oct4 维持胚胎干细胞自我更新状态的机制，诱导多能干细胞机制，肿瘤干细胞标志甚至肿瘤发生机制等重大科学问题无疑会有巨大的推动作用。

Oct4 的 $POU_{HD}$结构域的溶解状态在 1995 年已被 Morita 等用 NMR 的方法测定出来，但是至今还没有 $POU_S$ 结构域的具体资料。而 POU 家族的原初成员 Oct1 的 $POU_S$ 及 $POU_{HD}$结构域均已在 1999 年被测出（图 7.5，图 7.6）。现已证实，Oct4 的 POU 结构域与 Oct1 有约 60%相似，并且 Oct1 与 Oct4 的相互作用具有协同特征。因此，目前一些针对 Oct4 及其辅因子协同调控下游靶基因表达的分子机制的研究，正是先分析 Oct1/辅因子/DNA 三元复合体的结构，再通过计算机同源模拟建立 Oct4/辅因子/DNA 三元复合体模型的方法进行。

图 7.5　Oct1-POU 结构域/OCA-B 肽/DNA 三元复合物结构示意图

图 7.6　Oct1-POU 结构域/Sox2/DNA 三元复合物结构示意图

## 7.6 蛋白质三维结构研究方法

目前蛋白质结构的研究方法，主要可分为两大类：一是利用实验的方法，包括以 X 射线衍射（X-ray diffraction）或用核磁共振（nuclear magnetic resonance，NMR）的方法建立蛋白质结构模型；另一则是利用计算机，依据现有的结构信息结合理论运算，对蛋白质进行未知结构预测。预测方法包括同源模拟法（homology modeling）、折叠辨识法（folding recognition）以及从头算法（*ab initio*）三种。

（1）X 射线衍射　利用光的衍射特性，以 X 射线照射形成晶体的蛋白质，产生衍射图谱，再对此平面图谱进行一连串的计算分析，取得晶体的像角（phase angle），最后建立蛋白质结构模型。X 射线衍射法是目前获得蛋白质三维结构的主要途径，但其缺点是约有 50％的蛋白质不易取得晶体，且蛋白质纯化过程也有难度，这就需要更进一步改进表达系统（expression system）及结晶方法，以加速蛋白质结构的解析。

（2）核磁共振　利用电子的自旋（spin）特性，以外加磁场改变其自旋方式，藉此获得蛋白质的电子图谱，根据取得的数值计算蛋白质结构。由于 NMR 解出来的结构是蛋白质在水溶液中的状态，属于动态模拟方式，因而可以得到 20～40 种模型，最后经过运算取得平均结构。由 NMR 的结果，可以模拟蛋白质在体内的动态结构。但由于共振仪及信号强度的限制，NMR 能解出结构的蛋白质分子量约在 10kDa，约为 100 个氨基酸，所以蛋白质的长度限制了该方法的应用，若能提升 NMR 的解析能力，蛋白质结构的取得可快速增加。

（3）同源模拟法　将目标蛋白序列与现存于蛋白质结构数据库中的已知序列进行比对，找出最好的三维结构模板（template），以此模板为模型将目标序列导入后，取得后者的三维结构模型，经分子动能（molecular dynamics）及能量最小化（energy minimization）运算，最后可获得目标序列的蛋白质结构。同源模拟法的缺点是当现存数据库中没有类似的结构序列或序列相似度较低（＜30％）时，就无法以此方法进行预测，即使强行预测，结果的可信度也不高。

（4）折叠辨识法　将目标序列分区段代入已知蛋白质结构序列相似的模板，观察两者在某一区域内的相似度，经能量最小化运算，找出符合要求的分子。该方法需耗费大量的计算资源，对于蛋白质的核心部分具有较高的准确度，而外围的准确度则较差。

（5）从头算法　根据分子动力学原理，考虑氨基酸和溶液的所有交互作用力，找出分子间最稳定的状态，由一级结构开始，计算出蛋白质的三级结构。该方法同样需要大量计算，且只能对短链的肽进行预测。

综上所述，以实验方法建立模型共同的缺点是太过耗时，且常受限于样本的制备技术。因此，在蛋白质尚未以实验方式取得结构前，常常进行结构模拟。目前较理想的方法是分析蛋白质序列，在现存蛋白质结构数据库中取得原子三维坐标，通过电脑运算构建蛋白质的结构模型，经结构最优化，评估得到的蛋白质结构，最后取得最优的三维结构模型。

常用蛋白质分析相关网站及地址见表 7.1。

**表 7.1　常用蛋白质分析相关网站及链接地址**

| 名　称 | 链 接 地 址 |
|---|---|
| 数据库 | |
| SRS | http://srs.ebi.ac.uk/ |
| GenBank | http://www.ncbi.nlm.nih.gov/Genbank/index.html |
| SWISS-PROT | http://kr.expasy.org/sprot/ |
| PDB | http://www.rcsb.org/pdb/ |
| SCOP | http://scop.mrc-lmb.cam.ac.uk/scop/ |
| CATH | http://www.biochem.ucl.ac.uk/bsm/cath_new/index.html |
| DALI | http://www.ebi.ac.uk/dali/ |
| 模板搜索 | |
| BLAST | http://www.ncbi.nlm.nih.gov/BLAST/ |
| Fasta3 | http://www.ebi.ac.uk/fasta3/ |
| FUGUE | http://www-cryst.bioc.cam.ac.uk/~fugue/prfsearch.html |
| 二级结构预测 | |
| PHD | http://www.embl-heidelberg.de/predictprotein/predictprotein.html |
| SOPMA | http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl? page= npsa_sopma.html |
| nnPredict | http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html |
| 指认(**Threading**) | |
| THREADER | http://insulin.brunel.ac.uk/threader/threader.html |
| TOPITS | http://www.embl-heidelberg.de/predictprotein/submit_exp.html |
| 123D+ | http://www-lmmb.ncifcrf.gov/~nicka/123D.html |
| FRSVR | http://www.doe-mbi.ucla.edu/~frsvr/frsvr.html |
| 3D-PSSM | http://www.bmm.icnet.uk/~3dpssm/ |
| 序列比对 | |
| ClustalW | http://www.ebi.ac.uk/clustalw/ |
| BCM server | http://searchlauncher.bcm.tmc.edu/ |
| 同源模拟 | |
| SWISS-MODEL | http://www.expasy.org/swissmod/SWISS-MODEL.html |
| MODELLER | http://guitar.rockefeller.edu/modeller/ |
| WHATIF | http://www.cmbi.kun.nl/whatif/ |
| InsightⅡ/QUANTA | http://www.msi.com/ |
| 模型验证 | |
| PROCHECK | http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html |
| BIOTECK | http://biotech.embl-heidelberg.de:8400/ |

续表

| 名　称 | 链 接 地 址 |
| --- | --- |
| Prosa Ⅱ | http://www.came.sbg.ac.at/Services/prosa.html |
| AQUA | http://www.nmr.chem.uu.nl/users/jurgen/Aqua/server/ |
| ERRAT | http://www.doe-mbi.ucla.edu/Services/ERRAT/ |
| SQUID | http://www.ysbl.york.ac.uk/~oldfield/squid/ |
| VERIFY3D | http://www.doe-mbi.ucla.edu/Services/Verify_3D/ |
| 显示工具 | |
| RasMol | http://www.umass.edu/microbio/rasmol/ |
| SWISS-PDBViewer | http://www.expasy.org/spdbv/ |
| Cn3D | http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml |

## 7.7 同源模拟法

同源模拟是蛋白质三维结构预测的主要手段，也是目前为止蛋白质结构预测中最为成功、准确度最高的方法。但同源模拟有一个前提条件，即要模拟的目标蛋白必须有一个或多个已知结构的与之序列高度相似的同源蛋白。通常情况下，目标蛋白与模拟蛋白的序列一致性不低于30%，如果二者能够进行联配的区域长度超过100个氨基酸残基，序列一致性在25%以上，也可以进行同源建模。如果模板蛋白与目标序列的序列一致性在15%～25%，即所谓的“交界区域”(twilight zone)，则不能利用简单的同源模拟来预测结构。

同源模拟过程比较复杂，可分为以下几个步骤：

① 寻找一个或一组与待测蛋白质同源的通过实验测定的蛋白质结构，进行结构叠和；

② 建立未知蛋白质与已知结构蛋白质的序列比对；

③ 找出结构保守性的主链结构片段；

④ 建模结构变化的区域，一般为环区；

⑤ 侧链建模；

⑥ 利用能量计算的方法进行结构优化；

⑦ 评估。

目前互联网上提供了很多利用同源模拟技术预测的蛋白质结构的数据库，比如SDSCIS (San Diego)、3D-JIGSAW (London)、FAMS (Tokyo)和SWISS-MODEL (Geneva)，其中Guex等人发展的SWISS-MODEL是使用最广泛的网络建模服务器。

SWISS-MODEL (http://swissmodel.expasy.org) 由Manuel Peitseh于1993年发起成立，在Glaxo Smith Kline Geneva (GSK)进一步发展，通

过 ExPASy 服务器的瑞士生物信息中心（Swiss Institute of Bioinformatics，SIB）为公众提供服务。SWISS-MODEL 是一个有注解的基于同源建模的蛋白质结构服务器，它与专家蛋白质分析系统（Expert Protein Analysis System，ExPASy）网站是紧密联系的，其中有的也来自于 DeepView（Swiss Pdb-Viewer）程序。此服务器的目的是向全世界的生物化学和分子生物学工作者提供蛋白质三维建模服务。它开创了自动建模的先河，是迄今为止应用最广泛的免费在线服务器之一。至 2008 年 9 月，SWISS-MODEL 数据库共收录数据 340 万条，覆盖了 UniProt 数据库中 270 万种不同的蛋白质序列。SWISS-MODEL 数据库保持定期更新，保证了收录数据的全面性和准确性。

SWISS-MODEL 的工作模式有三种：自动模式（Automated Mode），比对模式（Alignment Mode）及项目模式（Project Mode）。

（1）自动模式　用户可直接通过 web 界面提交氨基酸序列或 Swiss-prot/TrEMBL 编号（accession），服务器会全自动为目标序列建立模型。如果一条模板与提交的目标序列相似度大于 25%，建模程序就会自动开始运行。但是，建模结果的可靠性会随着模板与目标序列之间相似度的降低而降低，如果相似度不到 50%，往往就需要用手工来调整序列比对。这种模式只能进行大于 25 个残基的单链蛋白质三维结构预测。

（2）比对模式　用户可以通过 SWISS-MODEL 比对界面上传 FASTA、MSF、CLUSTALW、PFAM 或 SELEX 格式的多重序列比对结果。当然，此序列中必须至少包含一条目标序列和一条来自于 Expasy Protein Database（ExPdb）的模板序列。

（3）项目模式　用户可将经过手工优化的请求提交给服务器。即可使用 DeepView 来建立一个项目文件，它包含模板结构、目标序列和模板序列的比对结果。

## 7.8 转录因子 Oct4 的 $POU_S$ 结构域同源模拟

本节通过计算机模拟的方法对 Oct4 的 $POU_S$ 结构域进行模拟预测。鉴于 Oct1 的 $POU_S$ 结构域的结构已在 1999 年通过 X-ray 方法获得，Oct4 的 POU 结构域与 Oct1 的有约 60%相似，因而完全可以采用同源模拟的方法对 Oct4 的 $POU_S$ 结构域进行预测。在此我们采用的同源模拟工具是 SWISS-MODEL 中的自动模式。

**【实验步骤】**

一、在 UniProt 数据库中获取 Oct4 的 $POU_S$ 结构域的序列

1. 访问 UniProt（http：//www. uniprot. orp/）数据库的 web 界面

（图 7.7）。

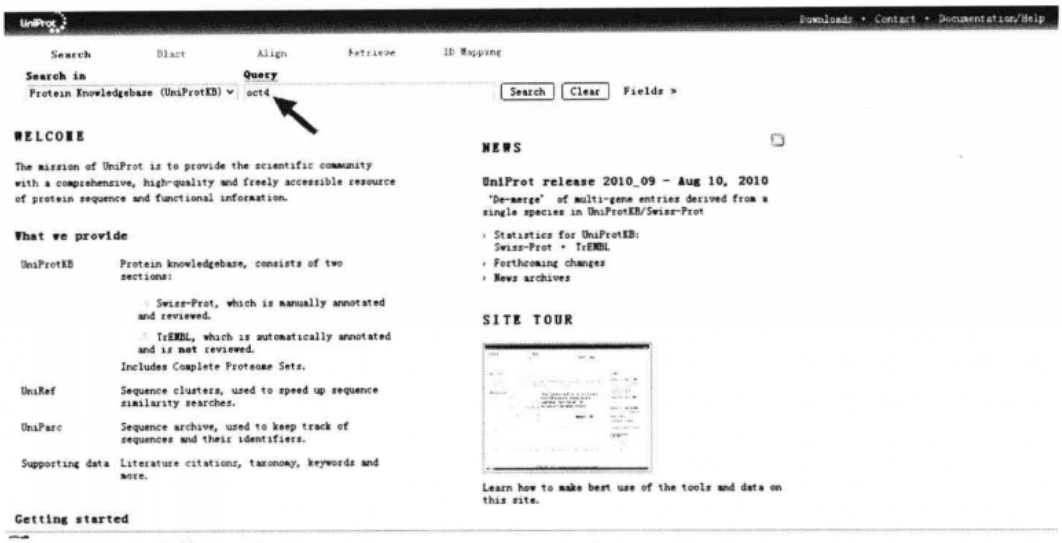



图 7.7　UniProt 数据库的初始界面

2. 在“Query”对话框中输入 Oct4，点击“Search”搜索该蛋白，数秒钟后显示搜索结果。结果包括鼠、人、牛、兔等不同种属的特异蛋白（图 7.8）。

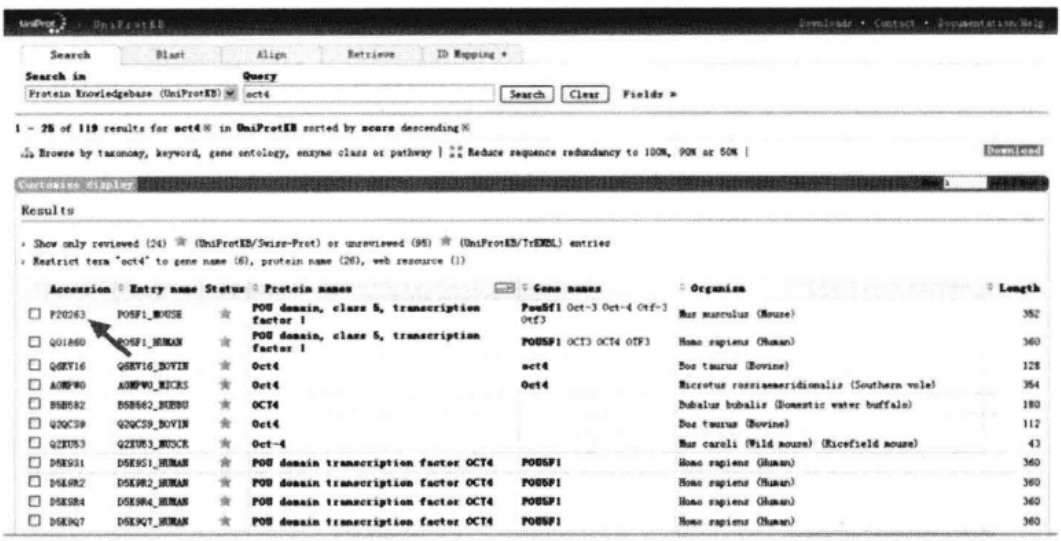



图 7.8　Oct4 蛋白在 UniProt 中的搜索结果

3. 选择分析鼠源 Oct4 蛋白，点击该蛋白的 Accession（P20263），查看其信息（图 7.9）。

4. 鼠源 Oct4 蛋白的 $POU_S$ 结构域由 75 个氨基酸组成，氨基酸编号为 131-205。点击这一段氨基酸，获取其序列（图 7-10）。

二、利用 SWISS-MODEL 中的自动模式进行同源建模

① 访问 SWISS-MODEL（http：//swissmodel. expassy. org/）服务器的 web 界面（图 7.11）。

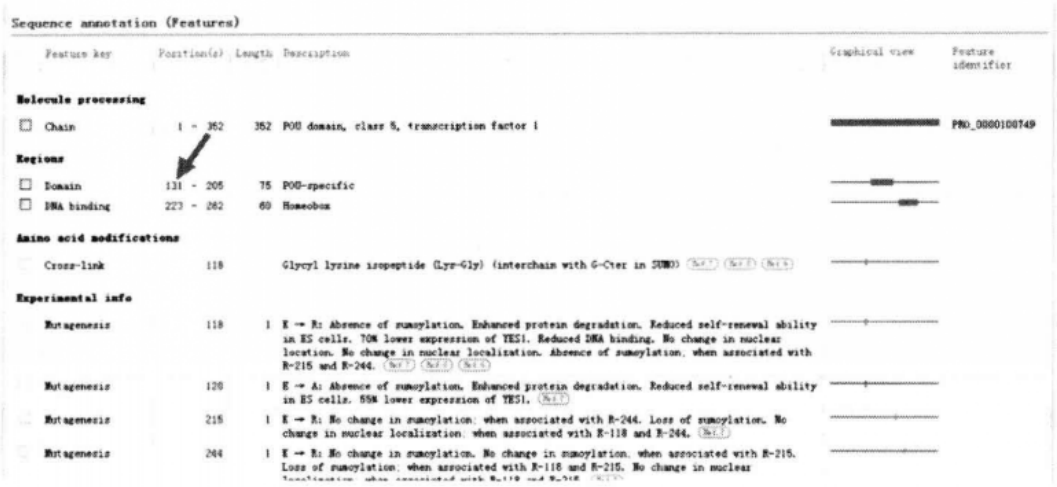



图 7.9　鼠源 Oct4 蛋白的信息

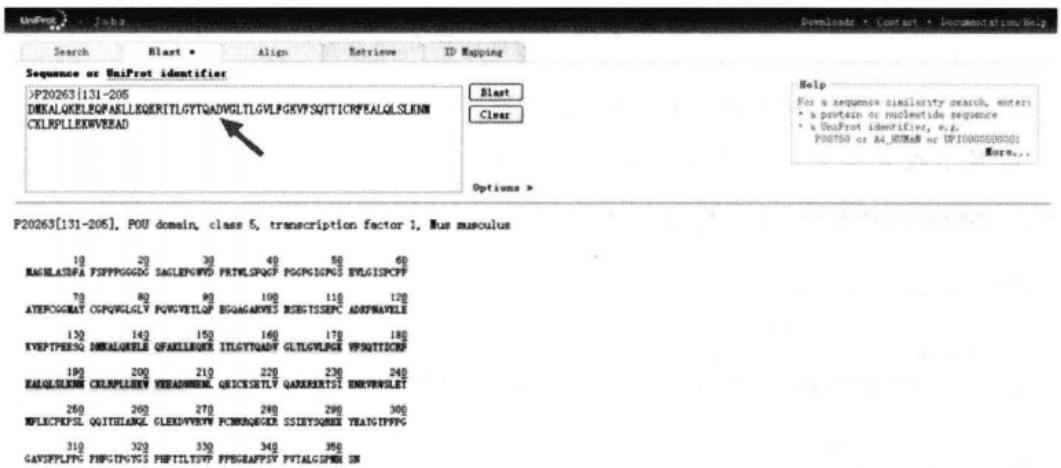



图 7.10　$POU_S$ 结构域序列

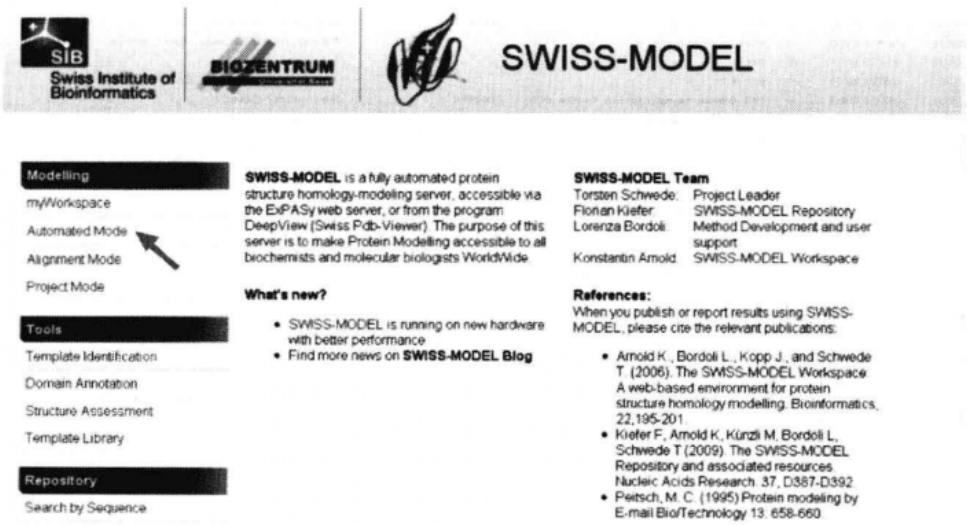



图 7.11　SWISS-MODEL 服务器初始界面

② 在“Modelling”一栏下，点击“自动模式”（Automated Mode），进入自动建模模式界面（图 7.12）。在“Email”对话框中填写自己的电子邮箱（电子邮箱可以填写也可以不填写，如果填写，提交任务完成后，会收到邮件提醒）；在“Project title”对话框中为本次建模操作填写一个名称，如 Oct4 POUs；在“Provide a protein sequence or a UniProt AC Code”对话框中输入通过步骤一获取的 POUs 结构域的序列；点击“Submit Modelling Request”提交任务。

图 7.12　SWISS-MODEL 服务器自动建模模式界面

③ 提交任务大约几分钟之后，邮箱会收到建模任务已经完成的邮件。同时 SWISS-MODEL 服务器页面也会自动更新，显示建模结果（图 7.13），结果可以 PDB 格式查看或者输出。

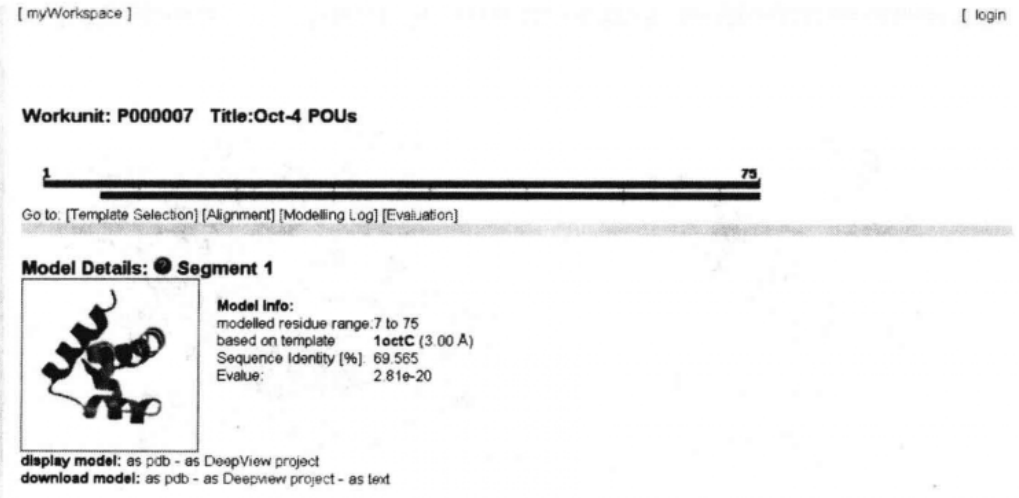



图 7.13　SWISS-MODEL 服务器的建模结果

**【实验结果】**

一、蛋白 P20263 的序列及相关信息

① 蛋白质全称：转录因子 Octamer-binding protein 4（Oct4）。

② 基因名称：Pou5f1。

③ 组织来源：小家鼠。

④ 序列长度：352 个氨基酸。

⑤ 存在阶段：出现在转录水平。

⑥ 蛋白质功能：与 SOX2 及 DNA 结合形成三元复合物，调控胚胎发育早期基因的表达。

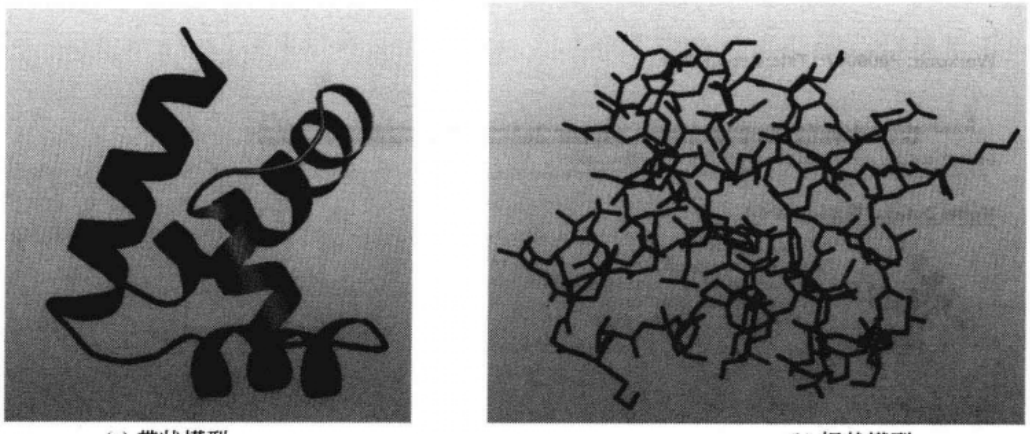
⑦ 其 $POU_S$ 结构域长度：75 个氨基酸。

⑧ 其 $POU_S$ 结构域序列：DMKALQKELEQFAKLLKQKRITLGYTQAD-VGLTLGVLFGKVFSQTTICRFEALQLSLKNMCKLRPLLEKWVEEAD。

**Alignment** [top]

```
TARGET     7         KELEQF AKLLKQKRIT LGYTQADVGL TLGVLFGKVF SQTTICRFEA
1octC      5     dl--eeleqf aktfkqrrik lgftqgdvgl amgklygndf sqttisrfea

TARGET               hhhhh hhhhhhhhhh      hhhhhh hhh           hhhhhhhh
1octC                hhhhhh hhhhhhhhhh     hhhhhh hhh           hhhhhhhh

TARGET     53    LQLSLKNMCK LRPLLEKWVE EAD ------ ---------- ----------
1octC      53    lnlsfknmck lkpllekwln daerkkrtsi etnirvalek sflenqkpts

TARGET               hhhhhh hhhhhhhhhh hh
1octC                hhhhhh hhhhhhhhhh hh          hhhhhhhhh hhhh     h

TARGET               ---------- ---------- ---------- ---
1octC      129   eeitmiadql nmekevirvw fcnrrqkekr inp

TARGET
1octC              hhhhhhhhh      hhhhhh hhhhhhhhhh h
```

图 7.14　鼠源 Oct4 蛋白 $POU_S$ 结构域与 1octC 的序列比对



(a) 带状模型　　(b) 棍状模型

图 7.15　鼠源 Oct4 蛋白 $POU_S$ 结构域三级结构

二、$POU_S$ 结构域三级结构建模

在 SWISS-MODEL 下用自动模式（automated mode）进行蛋白质三级结构同源建模。

找到的最匹配的模板为：1octC（3.00Å）。

序列的匹配度为 69.569%。

评价参数：2.81e-20。

目标蛋白与模板序列比对如图 7.14。

目标蛋白三级结构模型如图 7.15。

## 参考文献


[1] Becker A J, McCulloch E A & Till J E. Nature, 1963, 197: 452-467.

[2] Siminovitch L, McCulloch E A & Till J E. J Cell Comp Physiol, 1963, 62: 327-369.

[3] Liu H & Zhu S. Cell Stem Cell, 2008, 3 (6): 587-590.

[4] Takahashi K & Yamanaka S. Cell, 2006, 126 (4): 663-676.

[5] Maherali N, et al. Cell Stem Cell, 2007, 1 (1): 55-70.

[6] Okita K, lchisaka T & Yamanaka S. Nature, 2007, 448 (7151): 313-317.

[7] Wernig M, et al. Nature, 2007, 448 (7151): 318-324.

[8] Takahashi K, et al. Cell, 2007, 131 (5): 861-872.

[9] Yu J, et al. Science, 2007, 318 (5858): 1917-1920.

[10] Dohoon Kim & Chun-Hyung Kim . Cell Stem Cell, 2009, 4 (6): 472-476.

[11] Li W, Wei W & Zhu S. C. Cell Stem Gell, 2009, 4 (1): 16-19.

[12] Esteban M A, et al. J Biol Chem, 2009, 284 (26): 17634-17640.

[13] Qin D, et al. Cell Res, 2007, 17: 959-962.

[14] Liao J, et al. Cell Stem Cell, 2009, 4 (1): 11-15.

[15] Zhao X Y, et al. Nature, 2009, 461 (7260): 86-90.

[16] Niwa H, et al. Cell, 2005, 123 (5): 917-929.

[17] Herr W & Cleary M A. Genes Dev, 1995, 9 (14): 1679-1693.

[18] Pesce M & Scholer H R. Stem Cells, 2001, 19: 271-278.

[19] Rao M. Dev Biol, 2004, 275 (2): 269-286.

[20] Tai M H, et al. Carcinogenesis, 2005, 26 (2): 495-502.

[21] Yu H, et al. Am J Pathol, 2006, 168 (6): 1879-1888.

[22] Tondreau T, et al. Stem Cells, 2005, 23 (8): 1105-1112.

[23] Okuda T, et al. Brain Res Mol Brain Res, 2004, 132 (1): 18-30.

[24] Hochedlinoer K, et al. Cell, 2005, 121 (3): 465-477.

[25] Atlasi Y, et al. Int J Cancer, 2007, 120 (7): 1598-1602.

[26] Ezeh UI, et al. Cancer, 2005, 104 (15): 2255-2265.

[27] Iki K & Pour P M. Pancreatology, 2006, 6 (4): 406-413.

[28] Gibbs C P, et al. Neoplasia, 2005, 7 (11): 967-976.

[29] Looijengal H, et al. Cancer Res, 2003, 63 (9): 2244-2250.

[30] Liedtke S, et al. Cell Stem Cell, 2007, 1 (4): 364-366.

[31] Cantz T, et al. Stem Cells, 2008, 26 (3): 692-697.

[32] Morita E H, et al. Protein Sci, 1995, 4 (4): 729-739.
[33] Chasman D, et al. Genes Dev, 1999, 13 (20): 2650-2657.
[34] Ambrosetti D C, Basilico C & Dailey L. Mol Cell Biol, 1997, 17: 6321-6329.
[35] Attila Remènyi, et al. Genes Dev, 2003, 17: 2048-2059.
[36] Tress M, Ezkurdia I & Grana O. Proteins, 2005, 61 (7): 27-45.
[37] Arnold K, et al. Bioinformatics, 2006, 22: 195-201.
[38] Kiefer F, et al. Nucleic Acids Research, 2009, 37: D387-D392.
[39] Peitsch M C. Bio/Technology, 1995, 13: 658-660.
[40] David Cyranoski. Nature, 2008, 452: 406-408.
[41] Jeong Beom Kim & Hans R Scholer. Cell, 2009, 136: 411-419.
[42] 岳俊杰，冯华，梁龙．蛋白质结构预测实验指南．北京：化学工业出版社，2010.
[43] 张雷鸣，张剑宁．中华神经外科疾病研究杂志，2009，1671-2897：470-472.
[44] 谌容等．生命的化学，2006，26 (1)：54-56.
[45] 李明等．生物技术，2008，1 (3)：87-89.

# 第8章

# 蛋白质（酶）进化工程

冯　雁

蛋白质（酶）进化工程就是应用分子生物学及相关技术手段，通过对天然蛋白质（酶）的体外进化改造，获得性能更适合工业需求的、人工进化的非天然蛋白质（酶）。酶作为具有催化功能的蛋白质，和化学催化剂相比，其催化反应条件温和（常温、常压和中性 pH 等），效率高（化学催化剂的 $10^3 \sim 10^5$ 倍），而且底物选择性强。酶制剂已应用到食品工业、医药业、环保、能源开发和生命科学等各个方面，随着节能减排等社会经济发展的强烈需求，酶应用领域不断拓展。在实际工业生产条件下（高温、高压、极端 pH，以及存在重金属离子和氧化剂等蛋白质变性剂），天然酶结构和功能常常遭到破坏，从而使酶应用受到极大限制，发展酶分子进化的理论和技术成为全球关注的焦点。

早在 1978 年，加拿大科学家 Michael Smith 利用蛋白质结构信息，建立了定点突变技术，不仅对蛋白质中特定氨基酸的作用进行了探讨，而且也在体外对蛋白质功能进行了进化和改造。Michael Smith 因为建立了在蛋白质结构分析基础上的定点突变技术，于 1993 年获得了诺贝尔奖。近年来，随着酶结构-功能关系以及蛋白质分子进化规律了解的深入，科学家们又发展了结构基序嫁接和结构域重组（domain swapping）。前者代表技术如北京大学化学工程学院来鲁华教授于 2000 年报道了一种蛋白质-蛋白质相互作用表面位点的嫁接方法，取得了不错的成果。首先确定被嫁接的配体表面结合区的重要残基，在骨架蛋白中搜索适合于嫁接这些残基的位点，基于对映残基的 $C\alpha$，$C\beta$ 原子坐标，将骨架蛋白叠加到配体蛋白上，评估骨架蛋白与受体蛋白的互补性，并留下互补性较好的骨架蛋白。细调骨架蛋白与受体蛋白的相对位置，使之具有合理的界面堆积密度，然后在骨架蛋白选出的位置上移植配体蛋白相应的残基。结构域重组指募集已进化的蛋白质结构模块，根据所需要的功能进

行重组，因此突破了对关键残基突变仅能对酶部分性质改变的局限性，在较大范围快速实现蛋白质功能的改造。但由于目前对酶分子结构及催化机理研究还有限，使用理性设计方法的成功率还不高；同时蛋白质功能的多样性表明，还有很多的序列和结构空间没有被探测到，例如，100 个氨基酸构成的很小的蛋白质，序列上每个位点都可能存在 20 种氨基酸的任意一种，最后所有可能的组合是 $20^{100}$，这是一个非常庞大的数目，留给科学家对蛋白质未知序列和结构的探索空间。

20 世纪 90 年代发明了蛋白质非合理设计技术-定向进化（directed evolution）技术，定向进化又称为蛋白质的体外分子进化，是蛋白质工程的新策略，是在试管中模拟达尔文进化的过程，在不了解蛋白质结构信息的情况下，利用分子生物学手段对蛋白质基因的序列进行突变，创造分子的多样性、结合灵敏的筛选技术，迅速得到理想的突变体。定向进化技术应用自然进化原理来发展非理性的蛋白质设计技术，将更有可能获得性能更加优良或全新的酶，这是对蛋白质理性设计的完美补充。美国加州理工学院化学工程和生物化学教授 Arnold 女士认为："定向进化是'物竞天择'的一个高级版本，是在分子水平上的进化，研究的对象不是整个机体，而是特定的基因或蛋白质。"她领导的研究小组是定向进化领域的权威实验室，从 20 世纪 80 年代末就开始研究定向进化，1993 年，Arnold 应用分子进化原理创造性地利用易错 PCR（error-prone PCR，ep-PCR）对蛋白质进行改造，其后，又发展和建立了一系列定向进化技术和方法。由于定向技术在阐明蛋白质分子进化过程和机制方面的贡献，Arnold 荣获了美国三院院士（科学院、工程院和医学院）称号，成为八个健在的美国三院院士之一，也是唯一的女性三院院士，并成为世界公认的定向进化、生物催化和蛋白质工程专家。她主编了蛋白质定向进化的两本"圣经"——定向进化建库方法和定向进化筛选方法。随后，美国 Affymax 研究所的 Stemmer 教授于 1994 年引入 DNA 改组技术到蛋白质的定向进化过程中，并在 1997 年以分子育种定向进化为技术平台建立了从事定向进化的生物制药公司 Maxygen，大量制造功能改进的蛋白药物产品。

随着生命科学的发展，十几年来，定向进化技术以强大的生命力，在世界各地实验室"生根发芽，开花结果"，F. Arnold 课题组 1998 年提出的交错延伸（stagger extension process，StEP）和体外随机引发重组（random priming in vitro recombination，RPR），宾夕法尼亚州立大学化学系 S. Benkovic 教授于 1999 年建立的渐进式切割产生杂合酶技术（incremental truncation for creation of hybrid enzymes，ITCHY），2001 年美国 Enchira 生物技术公司的 Coco 博士发明的过渡模板随机嵌合生长重组技术（random chimeragenesis on transient templates，RACHITT），以及

德国不莱梅国际大学 Schwaneberg 课题组于 2004 年提出的序列饱和突变技术（sequence saturation mutagenesis，SeSaM）等方法，蛋白质定向进化技术不断走向成熟，为其他研究者提供了很多有价值的经验和可借鉴的方法。

尽管理性设计和定向进化在蛋白质设计及功能改造方面取得了巨大成功，但两者却各有优劣。理性设计在分子内引入定点突变，对关键残基或结构元素进行改变，但往往受限于蛋白质结构和功能间关系的复杂性，并且目前对酶结构解析数量还有限，使用理性设计方法的成功率还不高；同时蛋白质功能的多样性表明，还有很大的序列和结构空间没有被探测到。而定向进化虽然能够克服蛋白质结构和功能关系的障碍，而且常常用于产生有利于蛋白质折叠和功能发挥的一些精细结构变化，但是需要建立一个针对目标性质的高通量筛选方法来实现，通常的高通量方法都会消耗大量的人力和物力。而且全基因范围的随机突变，往往会造成人力、物力的浪费，蛋白质序列的长度决定了不可能将蛋白质序列上所有的突变组合都包含在内，因为这样会超过绝大多数筛选方法的承受范围。

近年来，基因组学、功能基因组学、结构生物学、生物信息学等多学科迅猛发展，酶结构-功能知识不断丰富，综合上述两种方法所发展起来的半合理设计（semi-rational design）结合了两种方法的优点，以蛋白质结构为基础，在选定的关键功能决定区域来进行随机突变，建立小型、智能型突变库，并筛选获得许多具有优良性状的重要突变体，有利于促进快速的蛋白质体外进化过程，极大地缩短了自然进化进程，从而发展和丰富蛋白质资源，这是蛋白质进化工程技术发展的一大进步。F. Arnold 研究组设计了 SCHEME 方法，即应用计算生物学技术对不同母体酶的进行结构解析，并在保持最小能量变化的基础上构建嵌合酶库，筛选获得了稳定性提高的纤维素酶；在半合理设计方法的探索研究中，德国科学家也作出了巨大的贡献。德国格赖夫斯瓦尔德大学 Bornscheuer 教授于 2008 年提出一种新的靶向定向进化方法——OSCARR 方法（one-pot simple methodology for cassette randomization and recombination）；德国马普学会煤化所所长 Reetz 教授提出了两种半理性设计方法——活性位点组合饱和突变方法（combinatorial active-site saturation test，CAST）和串联饱和突变方法（iterative saturation mutagenesis，ISM）。

蛋白质（酶）进化工程技术的不断发展，使人们对蛋白质功能设计和体外进化的能力有了重要提升，经过不断实践，获得了多种适于医药工业、化工、能源等领域应用的生物催化剂。下面就具体介绍各种蛋白质进化技术所取得的重要成果。

# 8.1 蛋白质的理性设计

蛋白质理性设计是基于对蛋白质结构的认识，通过定点突变技术改造蛋白质功能的一种方法。利用酶理性设计获得具有理想性质以适用于工业应用的突变体，关键是确定突变靶标。这就意味着在突变之前，人们需要做一个经过仔细考虑的猜测，在某一特定位点进行突变，获得所需性质的改变，这个方法通常要求有一个已经解析的蛋白质晶体。当然如果可以，最好同时获得酶与底物类似物或抑制剂的共结晶产物，这有助于更清晰地了解酶和底物的作用界面。随着计算生物学的发展，即使没有某一蛋白质解析的晶体结构，人们仍然可以通过计算机模拟构建结构模型，推测其可能的结构，获得其可能的结构。

## 8.1.1 蛋白质原有性质的改进

确定影响酶性质的关键氨基酸以及催化活性分歧的机制对进行酶理性设计是很有指导意义的，而确定导致酶性质不同的关键残基就是对酶家族进行序列和结构比对、自身三级结构和酶-配体复合物的分析。此外，由于酶的各种性质均与能量变化相关，而计算机模拟可以用来研究酶的活性中心、底物结合、柔性和解折叠，所以分子对接以及分子动力学模拟对于指导突变进行理性设计是很有意义的。

Kazlauskas 对以往酶性质改造的研究进行总结后，指出在多数情况下对活性中心较近的残基进行突变优于对远距离残基的突变，具体表现为对于立体选择性、底物特异性以及新的催化活性，近距离的残基突变要优于远距离残基突变；然而近距离和远距离突变均能改进活力、热稳定性以及对有机溶剂的稳定性。这也说明近距离的残基突变更有可能得到酶性质优越的突变体，揭示理性设计应该集中在对近距离的残基进行突变。

### 8.1.1.1 底物特异性与立体选择性的改造

对与底物相关性质进行理性改造，是建立在对酶和底物相互作用理解分析的基础上，与底物结合相关的关键氨基酸即组成底物进入通道，底物结合口袋的残基常常作为突变的靶标。

由于氨基转移酶具有广泛的底物谱、高立体选择性，以及不需要外部辅因子再生的特点，因此是很有前景的、能用于酶法合成光学活性胺的系统。来自 *Vibrio fluvialis* 的ω-氨基转移酶（ω-ATVf）对手性胺具有高的立体选择性，以及对芳香族胺出众的底物特异性，然而窄的底物特异性限制了其在脂肪族胺动力学拆分中的应用。Cho 等使用同源模建和底物-结构-活性关系

再设计的底物特异性，组成底物结合位点的大侧链残基 W57 和 W147 可能限制 ω-ATVf 对脂肪族胺的识别，突变体 W57G 和 W147G 的底物特异性发生改变，即在不影响对芳香胺水解活性和立体选择性的前提下，可以催化脂肪族胺的氨基转移。

微生物源 D-乙内酰脲酶联合 N-carbamoylase 可以用于合成光学纯度的 D 型和 L 型氨基酸，它们是合成半合成性抗生素、多肽激素、除虫菊素和杀虫剂的中间产物。尽管来源于 *Bacillus stearothermophilus* 的 D-乙内酰脲酶具有很多优良性质，但是其偏好于非取代的乙内酰脲，故利用结构和催化机制信息确定活性中心处 loop 区作为突变靶标，改变靶标氨基酸的侧链大小和疏水性强弱，突变体 M63I/F159S 对底物对羟基苯海因的特异性提高了 200 倍。

特定立体构型的羧酸酯是合成药物、杀虫剂和天然化合物的重要手性模块，很多具有立体选择性的脂肪酶仍然不能用于要求高纯度的药物工业应用。集中突变 *Yarrowia lipolytica* 脂肪酶底物结合位点处的五个残基（T88，V94，D97，V232，V285），V232A 立体选择性提高 10 倍左右，V232L 的立体选择性发生反转。以 *Burkholderia cepacia* 脂肪酶（BCL）底物结合口袋的三个残基（L17，V266 和 L287）作为突变靶标，得到 $E$ 值提高 10 倍的突变体 L17M 和 $E$ 值（对映选择性参数）增加 14 倍同时比活提高 15 倍的突变体 L17S/L287I。

#### 8.1.1.2 催化活力的改造

与底物立体选择性相比，影响催化活性的分子因素更容易预测，通过对底物与酶活性中心的相互作用以及活性中心结构特点进行分析，可以找到突破口。

基于计算机模拟（*in silico*）分子对接得到的酶-底物复合体可以指导确定抑制效果或活力提高的突变靶标。利用 *in silico* 分子对接分析谷胱甘肽转移酶（GST），将与外源物（除草剂和杀虫剂）结合相关的两个残基（Ile218 和 Gln53）作为靶标，Gln53Ala 对杀虫剂外源物的敏感度提高 9.2 倍。磷酸三酯酶（PTE）能够有效水解一系列有机磷（OP）杀虫剂从而达到解毒效果，故 PTE 可以作为游离酶生物治疗而商业化，并在动物研究中用于处理有机磷中毒。然而 PTE 并不能水解所有的有机磷，*Agrobacterium radiobacter* 磷酸三酯酶的底物对接显示两个残基（F131 和 W132）阻碍底物结合，突变体 W131H/F132A 对底物 Z-CVP 和 E-CVP 的水解活性分别提高 480 倍和 8 倍。

为了增加吡喃-2-氧化酶在生物燃料细胞和食品业的应用，对其活性中心残基进行突变，不仅得到了活力提高的突变体 V546C 和 V546C/E542K，

也得到热稳定性和活力同时提高的突变体 L537G 和 L537W。

漆酶（laccase）可以催化氧化一系列有机物，典型的就是酚底物，删除突变也是理性设计的一个重要手段，为了改进漆酶，应用定点突变和删除突变对其进行改造，得到酶活力和热稳定性均提高的突变体。

此外，理性设计也可以降低酶的副反应，嗜热 β-糖苷酶突变体 N282T 不仅具有高的转糖苷活性，且对供体邻硝基苯基 β-D-半乳糖苷（*o*NPGal）表现出与转糖活性竞争和产生副反应的自缩合活性，对亚位点［+1］/［+2］进行理性设计，以降低对 *o*NPGal 的结合力，结果 A221W/N282T 虽仍具有自缩合活性，但与 N282T 相比降低很多，而其转糖活性却几乎不变。

#### 8.1.1.3 稳定性改造

酶能否广泛应用于工业，也受热稳定性、有机稳定性以及 pH 耐受性等性质的影响。

最近证明利用计算机方法对蛋白质表面电荷-电荷之间相互作用能量的优化来理性设计提高热稳定性，同时保持活力不受损失，通过对不同大小和结构的两个酶酰基磷酸酶 AcPh 和 Cdc42 GTPase 进行改造，得到的突变体稳定性均优于野生型。同样基于计算模型理性设计 *Thermotoga maritima* 木聚糖酶 B 的 pH，与预期一致，对五个残基的取代（E70Q，E74Q，E77Q，G85Q 和 T87Q）改变了最适 pH，最适 pH 从 5.1 变为 5.5，活性 pH 范围也向碱性移动 0.5 个单位。

在体内，L-阿拉伯糖异构酶转化 L-阿拉伯糖为 L-核酮糖；在体外，也作为 D-塔格糖异构酶将 D-半乳糖异构化为塔格糖，高温和碱性 pH 下的异构化将引发副反应，而大部分 L-阿拉伯糖异构酶的最适 pH 处于碱性区，为了提高塔格糖的产率，对阿拉伯糖异构酶进行理性设计，Q268K 的耐酸能力更强，在酸性条件下的活力也增加；N175H 在较低温度 50～65℃范围内具有最高的活性，双点突变体 Q268K/N175H 综合了前二者的优点。

### 8.1.2 蛋白质优良性质的组合

将具有不同优越性质的同一家族酶进行结构域嫁接可能获得综合性质优越的突变体，从而满足工业应用的需求。基于 $(\beta/\alpha)_8$ 折叠家族 loop 嫁接的研究揭示，在同一家族中进行结构元件嫁接获得性质改进的突变体是可行的。

脂肪酸酰胺水解酶（FAAH）能水解大麻素和酰胺化的信号脂质，遗传学和药理学灭活的 FAAH 可以止痛，抗焦虑和消炎，表明 FAAH 将成为非常有前景的治疗靶标。然而受到人源 FAAH 表达的限制，利用分子生物学技术将人和鼠源 FAAH 的活性中心互换，得到的人/鼠 FAAH 嵌合体具有人 FAAH 的敏感性和鼠 FAAH 的高表达量。

由不同最适温度的同源酶产生嵌合体是获得理想最适温度酶的一种方法。Kamondi 使用来自 *Thermotoga maritima* 的两个纤维素酶：thermophilic xylanase A catalytic domain（TmxAcat，$T_{opt}=68℃$）和 hyperthermophilic xylanase B（TmxB，$T_{opt}=102℃$）通过理性的家族重组（rational family shuffling）构建了 21 个嵌合体，其中 4 个嵌合体的最适温度比 TmxAcat 的高，其中最稳定的嵌合体（$T_{opt}=80℃$）两个末端片段均来自 TmxB。进一步分析表明，N 和 C 末端片段之间的相互作用对整体热稳定性贡献很大；这对那些 N 末端和 C 末端相互联系的酶进行再设计具有一定的指导意义。

大部分脂肪酶具有一个盖子（lid）结构域来调节活性中心是否暴露，但是 *Bacillus subtilis* 脂肪酶 A（LipA）的活性中心却直接暴露于溶剂中，为了提高 LipA 对 1,2-*O*-isopropylidene-sn-glycerol（IPG）酯的动力学拆分，用 *Fusarium solani* 角质酶和 *Penicillium purpurogenum* 乙酰木聚糖酯酶较长 loop 取代 LipA 活性位点入口处的 loop，产生的杂合体对理想的 IPG 对映体表现了相反的立体选择性。

### 8.1.3 蛋白质新催化活性的获得

在自然界进化中得到的新酶相较于已有酶来说仅仅发生了微小的改变，这说明在现有酶基本骨架上只需要进行 loop 区域的插入、删除、替换或者氨基酸的取代，即可获得新酶活性，因此来自序列和结构方面的信息将指导活性转换。

由于苯甲酸酯水解酶在工业和医药业应用具有很大的潜力，Höst 期望使人碳酸酐酶（human carbonic anhydrase）Ⅱ具有对对硝基苯酚苯甲酸酯的水解活性，基于一个过渡态类似物和人碳酸酐酶Ⅱ活性中心的分子对接分析，成功构建了具有苯甲酸酯水解活性的三点突变体 V121A/V143A/T200A，它不仅对苯甲酸酯具有水解活性[$k_{cat}/K_m=625(\pm38)$ L/(mol·s)]，而且对其他酯底物也有很高的水解活性[对硝基苯酚乙酸酯的 $k_{cat}/K_m=101700(\pm4800)$ L/(mol·s)]，从而由碳酸酐酶获得了具有水解活性的酯酶。

Kazlauskas 和 Bornscheuer 两课题组期望将假单胞荧光杆菌酯酶（PFE）转化为环氧化物水解酶，通过对二者催化机制以及氨基酸序列的比较，确定了包括活性中心 Ser 在内的 8 个残基可能导致活性歧化，然而，构建的 7 个突变体都未表现出环氧化物水解酶活性。在分析了两种水解酶的结构后发现一段 loop 区对二者的功能差异可能起重要作用，将其嫁接最终成功获得具有环氧化物水解酶活性的突变体。这表明对基本骨架进行微调是可以获得新酶活性的，前提是不仅要了解催化机制和进行序列比对，更重要的是对靶标酶的三级结构进行分析。

对酶基本骨架微调可获得新活性，而上面的实例显示微调处也需要构象优化，以便功能残基的侧链精确定位，从而可以和新底物相互作用。基于这一理论，David Baker 课题组发明了一种计算机方法对活性中心附近 loop 区重建和再设计，以引入特异性的酶-底物相互作用，从而获得新酶活性。对人鸟嘌呤脱氨酶的关键 loop 区进行优化再设计，以产生关键的相互作用与底物 ammelide，设计的突变酶对 ammelide 的活力提高 100 倍。

与非金属酶相比而言，金属酶的理性设计更加富有挑战性，这不仅仅表现在金属结合位点多变和不同的金属离子氧化状态，而且计算机模拟时的力场也不确定。但是研究者仍然通过在肌红蛋白远端口袋引入 3 个 His 和 1 个 Glu，成功地设计出了一个金属蛋白的结构和功能模型——一氧化氮还原酶 NOR。

尽管合理设计已取得一定的成功，但是对很多蛋白质工程研究者来讲，这个方法仍然充满意外和艰辛，对蛋白质结构上的认知一旦出现偏差，就不会获得功能上的理想改变，而且就算是在对蛋白质结构和功能关系有着正确理解下的设计，也很难获得性质显著改善的突变体，而定向进化的出现很好地改变了这一状况。

## 8.2 蛋白质的定向进化

定向进化属于蛋白质的非理性设计，它不用事先了解蛋白质的结构、保守位点、催化机制等因素，而是人为地创造特殊的进化条件，模拟自然进化机制，在体外对酶基因进行改造，并定向筛选，获得具有某些预期特征的进化酶。酶定向进化技术在一定程度上弥补了定点突变技术的不足，具有很大的应用价值。

### 8.2.1 定向进化的原理及目的

定向进化是利用 Taq DNA 聚合酶不具有 3′→5′校对功能的性质，并结合基因工程现有技术手段，在待进化酶基因的 PCR 扩增反应中，配合适当条件，以很低的比率向目的基因中随机引入突变，并构建突变库，凭借定向的选择方法，选出所需性质优化的酶。

定向进化的基本规则是："获取所所需性质的突变体"。简言之，定向进化就是随机突变同筛选或选择相结合。与自然进化不同，定向进化是人为引发的，制造某一方向的进化而排除其他方向突变的作用，整个进化过程完全是在人的控制下，所以，定向进化又被称为"试管内进化"。酶分子定向进化的目的在于人为地改变天然生物催化剂的某些性质，增强稳定性，创造天然生物催化剂所不具备的某些优良特性，甚至创造出新的活性和产生新的催

化能力，扩大生物催化剂的应用范围。

### 8.2.2 定向进化的应用

定向进化在其出现的短短十几年内已发展成一门相当成熟的蛋白质及多肽的改造技术，它的应用极大地促进了酶工程、代谢工程以及医药等其他相关领域的发展。在众多研究者的共同努力下，蛋白质定向进化领域取得了显著成绩，定向进化应用于工业生物催化剂改造的成功例子已经很多。

(1) 提高酶分子的催化活力　苯甲酰甲酸脱羧酶的副反应是催化碳碳键形成的聚醛作用，来源于 *Pseudomonas putida* 的苯甲酰甲酸脱羧酶，通过定向进化获得了突变体 L476N，该突变体相对于野生型来讲其聚醛酶活力提高 5 倍，而且作者认为其 Leu476 位氨基酸为通过控制产物释放活性中心的“门户守卫者”。三磷酸酯酶通过基于随机突变的定向进化结合微珠法筛选，获得 63 倍于野生型活力提高的突变体。利用 DNA 改组方法，Stemmer 对 $\beta$-内酰胺酶进行了改造，以向培养物中添加头孢氨噻为选择压力，经三轮 DNA 改组和筛选，得到了酶活力提高 32000 倍的突变体。

(2) 提高酶分子的稳定性　Zhao H. 等利用连续的易错 PCR 获得数株阳性克隆，并用交错延伸方法重组 DNA，采用连续提高培养温度对突变体进行筛选，最终获得一株突变体，不仅将枯草杆菌蛋白酶的最适温度提高了 17℃，还使酶在 65℃的半衰期增加了 200 倍。定向进化也用于探查酶分子在进化过程中的热稳定性和酶活力之间的关系，Giver L. 等利用枯草杆菌来源的对硝基苯酚酯酶来研究进化过程中的热稳定变化，他们通过连续 4 轮随机突变和 1 轮 DNA 改组，获得了突变酶，$T_m$ 增加了 14℃。在 30℃时同时对酶活力和热稳定性进行筛选，获得了热稳定性提高，同时维持原酶活力的突变酶分子。Komeda 对来源于 *Ochrobactrum anthropi* SV3 的 D 型氨基酸酰胺酶的稳定性和活力进行了随机突变和筛选。突变克隆菌株在平板上原位裂解热处理，同时显色分析，获得一个最适反应温度提高 5℃且 $v_{max}$ 提高 3 倍的突变体。

(3) 提高酶分子的底物专一性　Zhang 等在定向进化实验中证明，对底物特异性影响的突变氨基酸不仅仅局限于底物的结合位点。他们利用定向进化把一个半乳糖苷酶改造成以果糖为底物的酶，虽然 $k_{cat}$ 下降 2 倍，但 $K_m$ 下降了 20 倍。突变体的序列分析表明，利用 7 轮进化筛选出的 6 个突变体中仅有 3 个位于底物结合位点附近，另外 3 个并不在底物结合口袋或酶活性位点附近。Alexeeva 等对来源于 *Aspergillus niger* 具有苄胺氧化作用的酶进化其对 $\alpha$-甲基苯胺作用能力，应用致错菌株 *E. coli* XL-1 red，通过平板显色法在含有 $\alpha$-甲基苯胺底物的平板上进行筛选，得到进化菌株 Asn336Ser 的对于 $\alpha$-甲基苯胺活力提高 50 倍的突变体。May 等对来源于 *Arthrobacter*

sp. DSM 的 D-海因酶进行了定向进化，使其对于 L-5-乙内酰脲的选择性增强，通过随机突变结合饱和突变的方法应用平板 pH 指示剂法得到具有所需性质的突变体。Reetz 和 Jaeger 等通过几轮的随机突变结合 DNA shuflling，对于脂肪酶的立体选择性进行了改进，经过 45000 个菌株的筛选获得了一个突变体，其对于 $R$ 型底物的 $E$ 值由 1.1 提高到 30。Williams 对催化立体化学的碳氮键形成的 1,6-二磷酸塔格糖醛醇酶的立体选择性采用 DNA shuflling 进行了定向进化，进化后的突变体的催化效率较野生型对非天然底物 1,6-二磷酸塔格糖提高了 80 倍。

随着体外基因多样性设计策略、高通量筛选技术的发展，以及应用范围的拓展，相信在不久的将来，人们能发现并创造出大量理想的蛋白质，提高其在实际应用中的使用价值。

## 8.3 蛋白质的半合理设计

虽然理性设计和定向进化在蛋白质设计上取得了巨大成功，但是两者作为蛋白质改造方法却各有优劣。理性设计一般用来在分子内引入随机突变不易获得的关键性的残基或结构元素，而定向进化常用来产生有利于蛋白质折叠和功能发挥的一些精细结构变化。因此，两者的结合将大大扩展蛋白质工程学的研究范围，制得具有许多优良性状的蛋白质，大大缩短自然进化的进程，从而极大地发展和丰富蛋白质资源，应当说这是蛋白质工程技术发展的一大进步。

此外，理性设计往往受限于蛋白质结构和功能间关系的复杂性，定向进化虽然能够克服蛋白质结构和功能关系的障碍，但是需要建立一个针对目标性质的高通量筛选方法，通常的高通量方法都会消耗大量的人力和物力。而且全基因范围的随机突变，往往会造成人力、物力的浪费，因为控制某一所需性质的位点，有可能仅仅局限于序列的某一部分。不仅如此，蛋白质序列的长度决定了不可能将蛋白质序列上所有的突变组合都包含在内，因为这样会超过绝大多数的筛选方法的承受范围。因此如果能够将两种方法结合，即半合理设计（semi-rational design），应用随机突变探测改变所需性质的关键位点，随后理性选定一些位点进行饱和突变，可提高对目标酶所需性质的改造效率。

### 8.3.1 基于结构的有选择的随机突变

一旦解析获得蛋白质的晶体结构，尤其是与底物类似物共结晶的结构，可以清楚地获得酶和底物的作用关系。但是仍然存在的问题是，作用位点可能是个很广泛的区域，并且对此区域随机突变的互相组合结果仍然是个只有

筛选能够解决的问题。对确定区域的随机突变，确实可以提高进化的效率，毕竟这相对于整个蛋白质序列来讲还是一个相对小并且更重要的范围。

Santoro 等应用这种针对目标序列的随机突变，来改变 Cre-DNA 重组酶的底物特异性，应用此酶与底物共结晶结构的信息，选择了两个有可能的短序列范围，分别包含 5 个和 6 个残基，建立随机突变库，获得一株原本野生型不具备的对新的底物序列识别的突变体，该方法可以很有效地进化到所需的性质。类似地，Rui 等对一种环氧化物水解酶的降解特性进行改造，也是采用选定的位点进行饱和突变，经过仔细的结构研究，最终在较小的库中挑选到一个活力提高 10 倍的突变体。

### 8.3.2 随机突变后的定点饱和突变

当结构信息不足以支持一个成功的合理设计时，可以采用随机突变作为先导，用来探测对所需性质的有益突变位点，再针对确定的有效位点施以饱和突变，可以很快速地提高所需的目标性质。Reetz 等采用随机突变结合筛选，对来源于 *Pseudomonas aeruginosa* 的酶对底物的立体选择性进行改造，四轮的随机突变使其 $E$ 值由 1.1 上升为 11.3，随后对随机突变中确定的“热点”（155 位和 162 位氨基酸）进行了饱和突变，获得了 $E$ 值达到 55 的更优突变体。Geddie 等对大肠杆菌的 β-葡萄糖苷酸酶的底物特异性进行半合理设计，针对几轮定向进化中获得“热点”的定点饱和突变，使其 β-樱草糖苷酶的活力再次提高近 70 倍，随后将优势突变体基因进行重组，使得催化活力提高至 100 倍。

### 8.3.3 随机突变与定点饱和突变同步

基于蛋白质结构的比对和序列比较，Peimbert 等在 *Streptococcus pneumonia* 转肽酶中引入 β-内酰胺酶活力，同步地采用全序列随机突变和选定位点的定点饱和突变。获得了一个具备较高 β-内酰胺酶活力的突变体。这个方法被证明有效地利用了已知的结构信息，同时采用随机突变探测未知的有效位点，同样可以获得所需性质的快速提高。

### 8.3.4 计算机辅助的半合理设计

通常的半合理设计可以很好地限定随机突变范围，使得随机突变更有目的性，但是尽管如此，选定可能范围后，最后的随机突变的库容量，仍然是很大的。事实上，对 5 个残基的 20 种氨基酸的随机突变，最后的库容是 $20^5 = 3.2 \times 10^6$。对通量不是很高的筛选方法，这个数量仍然庞大。借助计算机的帮助，采用模拟的方法，现在已经能够剔除不必要的突变，保证在尽可能小的库中，产生更多的有益突变。

Hayes 等用一种称为蛋白质设计自动化（protein design automation，PDA）的计算机算法，来预测能够形成正常蛋白质折叠的突变体，最后结果证明，这个方法剔除有害突变后，需要筛选的库容量下降了几个数量级。TEM-1 β-内酰胺酶的 19 个预先确定的残基随机突变库（$7\times10^{23}$）进行计算机模拟筛选，选定其中 200000 个能量最低结构的突变体构建突变库，在这个比原来小得多的突变库中获得了 1280 倍于野生型活力的突变体。

总的说来，半合理设计综合了合理设计和非合理设计两个方法的优点。针对自身实验材料的特点，灵活地将两种方法互相补充，可以提高对所需性质进行改良的速度。随着计算机模拟的生物信息学的不断发展，引入更加依赖于高性能、高精度、低成本的模拟计算可以为酶性质的改造提供更强的辅助。通过半合理设计，在对酶的改造上，人们可以采用更多更灵活的方法，制备出更优品质的酶，同时进一步加深对酶的结构和功能关系的理解。

## 8.4 应用实例——超嗜热酰基肽水解酶/酯酶的半理性设计

### 8.4.1 引言

在冯雁教授实验室的前期工作中，我们已克隆表达了来源于超嗜热古细菌 *Aeropyrum pernix* K1 的酰基肽水解酶基因 APE1547，酶学性质分析表明，所获得的重组蛋白 apAPH 主要表现出酰基肽水解酶和酯酶的双活性。对其晶体结构进行解析发现，该酶属于丝氨酸蛋白酶家族中新发现的脯氨酸寡肽酶（POP）亚家族，由一个 α/β 水解酶结构域和一个 β 推进器结构域组成，酶活性中心位于两个结构域之间。在获得酶结构信息的基础上，我们期望能阐明 apAPH 具有催化双功能性的结构基础，并通过定向进化获得酯酶活性提高的突变库。本节先通过序列及结构比对确定可能对酶底物特异性具有重要作用的 Arg526 位点，对其建立定点饱和突变体，并通过高通量筛选获得酯酶活力明显提高、肽酶活力明显下降的突变体，使 apAPH 由双功能酶变为具有良好底物专一性的酯酶。在此基础上，我们选择了具有代表性的突变体，对其进行了稳态动力学测定及分子动力学模拟，证明了 526 位点决定底物选择性的重要作用，并揭示了丝氨酸蛋白酶家族与脂肪酶家族的进化相关性。

### 8.4.2 实验方法

#### 8.4.2.1 同源序列比对

分别选取来源于 *Aeropyrum pernix* K1，*Caenorhabditis elegans*，鼠肝（rat liver），人红细胞（human erythrocytes），*Bacillus subtilis*，*Fugu ru-*

bribes 的酰基氨酰肽酶；来源于 *Pyrococcus horikoshii*，猪脑（porcine brain），牛脑（bovine brain），*Arabidopsis thaliana* 的脯氨酸寡肽酶；来源于 *Archaeoglobus fulgidus*，*Bacillus acidocaldariu*，*Alicyclobacillus acidocaldarius*，*Escherichia coli*，*Moraxella* 的激素敏感酯酶（hormone sensitive lipase，HSL）的序列中对应 apAPH 的 Arg526 位点附近的五肽氨基酸序列进行同源比较。

#### 8.4.2.2 蛋白质三级结构比对

采用混合延伸（combinatorial extension，CE）方法搜索蛋白质结构相似性，其算法是将蛋白质的结构元件作为可比对的结构片段（aligned fragment pairs），使用类似于序列同源性的算法得到结构相似程度高的蛋白质。CE 算法更多地基于片段相似性，而不是整体的拓扑结构。对搜索到的结构进行 RMSD（Root Mean Square Deviation）分析以定量其整体结构相似性。

RMSD 的计算公式如下：

$$\mathrm{RMSD}=\sqrt{\frac{\sum_i d_i^2}{n}}$$

式中，$n$ 为原子数；$d_i$ 为相应的 $i$ 原子间的距离。

结构相似的判定标准为：相同的结构（identical structures），其 RMSD=0；相似的结构（similar structures），其 RMSD=1～3Å；进化关系较远的结构（distant structures），其 RMSD>3Å。

#### 8.4.2.3 全质粒 PCR 构建定点饱和突变库

应用简易全质粒 PCR 突变法（QuickChange® site-directed mutagenesis protocol）建立 Arg526 位点的饱和突变库，引物设计如下：

R526N_f：5′-ATGAACGACAGCNNKACACCGCTGAGCATCGACCG-3′

R526N_r：5′-CGGTCGATGCTCAGCGGTGTMNNGCTCGTCTTCAT-3′

使用全质粒 PCR 反应构建饱和突变库，PCR 体系为：

| | | | |
|---|---|---|---|
| 野生型 pET 15b-apAPH | | dNTP | 0.2 mmol/L |
| 质粒模板 DNA | 50ng | DNA polymerase | 5 U |
| 上游引物 | 40 pmol | 10×Reaction buffer | 5 μL |
| 下游引物 | 40 pmol | $ddH_2O$ | 补充至 100 μL |

PCR 反应条件：94℃，1min；45℃，1min；72℃，10min，20 个循环；72℃，20min。反应结束后加入 2U DpnI 37℃保温 1h，消化野生型基因的质粒模板，利用 PCR 纯化试剂盒对 PCR 产物进行纯化，以除去盐离子及蛋白质，取 2μL 电转化入大肠杆菌 BL21（DE3）CodonPlus 感受态细胞。

#### 8.4.2.4 饱和突变库的筛选

使用实验室前期建立的针对嗜热酯酶的高通量筛选方法，对 apAPH 的饱和突变库进行了筛选，具体操作如图 8.1 所示。

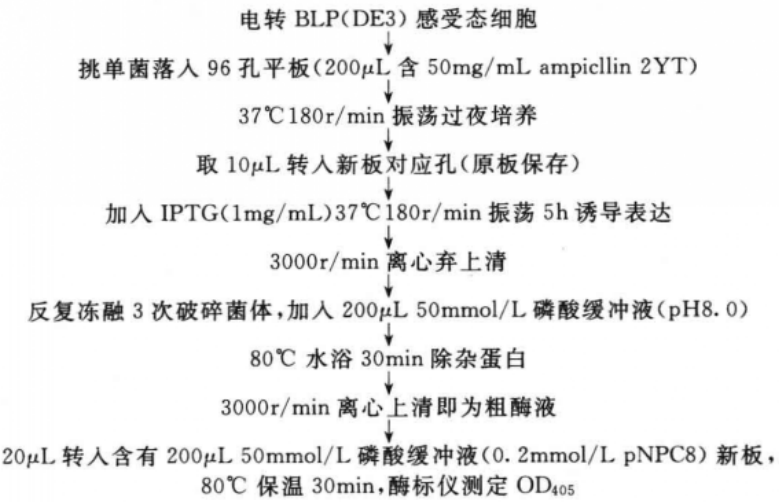



图 8.1 饱和突变库的筛选流程

#### 8.4.2.5 野生型 apAPH 及其突变体的表达及粗酶获得

选择 pET15b 作为质粒表达载体，*E. coli* BL21（DE3）CondonPlus 作为宿主细胞进行 apAPH 及其突变体的高效表达。

将野生型 apAPH 与其突变体的工程菌按 2%接种量接种于 5mL 含有 50μg/mL 氨苄青霉素的 2×YT 培养基中，在 37℃下振荡培养过夜。按同样接种量接种于 200mL 2×YT 培养基，37℃振荡培养 3～4h，然后接种于 2L 2×YT 培养基中，37℃振荡培养，当培养液的 600nm 吸光度值达到 0.6～1.0 时，加入诱导剂 IPTG 至终浓度为 1mmol/L，25℃培养过夜。在 6000r/min 下离心 20min，收集菌体，−20℃冷冻保存。

在室温下融化菌体，加入 5～10 倍体积的 50mmol/L 磷酸缓冲液(pH8.0)，超声波破碎细胞 10min，细胞破碎液于 85℃水浴保温半小时失活大部分杂蛋白，12000r/min 离心 20min 后收集上清，即得粗酶液。

#### 8.4.2.6 野生型 apAPH 及其突变体的镍亲和柱纯化

3mL 镍亲和柱（Ni-chelating column）加入去离子水至柱顶端，自然洗脱后用 5 倍体积的 1×Charge buffer 自然洗脱，再加入去离子水至柱顶端自然洗脱，后用 3 倍柱体积 1×Binding buffer 自然流干，再将 0.22μm 滤膜过滤的粗酶液上柱，待自然流干后用 10 倍柱体积的 Binding buffer 自然洗脱。最后用 6 倍柱体积的 1×Wash buffer 洗脱目的蛋白质，然后用 20mmol/L 磷酸缓冲液（pH8.0）透析蛋白质溶液除去咪唑，用聚乙二醇包埋浓缩目的蛋白，电泳检测纯度。

#### 8.4.2.7 酯酶、肽酶活力测定

以对硝基苯酚辛酸酯（*p*NP-C8）作为酯酶底物，缓冲体系为 50mmol/L 磷酸缓冲液（pH8.0），底物终浓度为 0.2mmol/L，反应体系为 1mL，酶适量。在 80℃用 2550 紫外可见分光光度计测定 405nm 的吸光度值的变化曲线。1 个酶活力单位定义为 1 分钟水解底物生成 1μmol 对硝基苯酚（$\varepsilon=0.016\text{L}/\mu\text{mol}$）所需要的酶量。酶活计算公式如下：

$$\text{酶活(U/mg)}=\frac{\Delta\text{OD}_{405}V}{t\varepsilon V_{\text{酶}}\ c_{\text{酶}}}$$

式中，$\Delta\text{OD}_{405}$为反应时间内 405nm 的吸光度变化值；$V$ 为反应体系体积（1mL）；ε 为对硝基苯酚的消光系数（0.016L/μmol）；$t$ 代表反应时间，min；$V_{\text{酶}}$ 是加入的酶量，mL；$c_{\text{酶}}$代表酶的浓度，mg/mL。

肽酶活力的测定以对硝基苯胺乙酰亮氨酸（Ac-Leu-*p*NA）为底物，对硝基苯酚的消光系数为（0.016L/μmol）。其余方法、数据处理手段等与酯酶活力测定相同。

#### 8.4.2.8 动力学常数测定

在 50mmol/L 磷酸缓冲液（pH 8.0），80℃条件下测定野生型 apAPH 及其突变体的活力，使用 *p*NPC8 和 Ac-Leu-*p*NA 作为底物测定其酯酶、肽酶活性，选定的浓度范围分别在 20～200μmol/L 和 200～2000μmol/L 之间。测定酶反应初速度并应用 Lineweaver-Burk 作图法对 1/*V*-1/[S] 作图，由在两轴上的截距求得 $K_{\text{m}}$ 值（$x$ 轴截距的负倒数）与 $V_{\text{max}}$（$y$ 轴截距的倒数），再由 $V_{\text{max}}=k_{\text{cat}}\times[\text{E}]$ 计算出 $k_{\text{cat}}$，从而得到动力学常数 $K_{\text{m}}$ 和 $k_{\text{cat}}$值。

#### 8.4.2.9 野生型 apAPH 及突变体底物复合物模型构建

所有的计算由 Insight Ⅱ 2000 软件完成（Accelrys，San Diego，CA），使用 BIOPOLYMER 模块用于小分子底物的构建，能量最小化和分子动力学模拟应用 DISCOVER 模块在连续共价力场下完成。使用 apAPH 1.8Å 晶体结构（PDB id：1VE6）作为计算初始蛋白质模版。氢原子按照 pH 8.0 时残基的解离状态添加到蛋白质分子中，原子电势按照连续共价立场力场内加入，在酶分子表面加添 5Å 水分子层。

小分子 Ac-Leu-*p*NA 与蛋白质的复合物结构采用手动对接方式生成，底物的亮氨酸侧链摆放在肽底物 S1 结合位点，同时保证其酰基氧位于 Arg526 残基胍基集团的附近，之后进行能量最小化，获得构象为后续突变体模型构建的初始模板。分子对接模板交由 PROCHECK 程序绘制拉氏构象图，检测其结构合理性。

使用高级分子对接程序 Affinity 对小分子 *p*NPC8 进行自动分子对接，

应用模拟退火和蒙特卡洛法，对 apAPH 和底物对接的最佳构象按照能量标准进行优化选择，最终选取能量较低且小分子位置摆放有利于反应进行的构象，将对接结果交由 PROCHECK 程序绘制拉氏构象图，检测其结构合理性。

在野生型 apAPH 与底物复合物的基础上，使用 BIOPOLYMER 对 Arg526 位残基进行取代，构建突变体的模型。保持替换残基侧链的方向与 Arg526 方向一致，并进行能量最小化，利用 PROCHECK 程序绘制拉氏构象图，检测其结构合理性。

#### 8.4.2.10 分子动力学模拟

分子动力学模拟（MDS）计算由 Insight Ⅱ 2000 软件完成。在前期构建的酶分子与底物结合的复合结构中，固定以活性中心 Ser445 为中心 15Å 范围外的残基，其余氨基酸与水分子的作用包括在计算之内。设定非共价作用距离为 20Å，在分子动力学模拟之前先使用陡降法进行 100 步能量最小化，之后应用梯度退火方法进行能量最小化，能量梯度收敛至 0.001 kcal/(mol·Å)。能量收敛体系在经过 100fs 的预平衡后，在 310K 进行 100ps 的动力学模拟，以 1fs 为一步，共收集 101 个构象，对每个构象中关键残基与底物的相对位置、相互作用进行仔细分析，解释酶动力学所获得的结果。

### 8.4.3 实验结果

#### 8.4.3.1 半理性设计思路的确立及突变位点选择

衡量蛋白质的进化关系最常用的方法是同源序列比对。然而，对于经历了长期趋异进化的蛋白质家族来说，利用三级结构重合比较（structure alignment）来确定蛋白质进化关系的方法可以提供很多序列同源性无法提供的信息。我们利用 CE 算法，对 apAPH 进行了结构相似性搜索，发现 apAPH 的催化结构域与激素敏感脂肪酶（HSL）家族具有惊人的相似性（图 8.2），来自嗜热古菌 *Archaeoglobus fulgidus* 的酯酶 AFEST 的主链 C 原子 RMSD 仅为 2.3Å。一般来源于共同祖先的蛋白质，在序列同源性低于 25%时，两者的主链重合偏差值仍然可能接近 2Å，而 apAPH 与 AFEST 的序列同源性仅为 18%，但是两者的主链重合偏差仅为 2.3Å，并且在底物选择上两者都偏嗜选择酰基链长度为 8 个碳原子左右的酯，这种结构和功能上的相似性表明两者在进化上密切相关，有可能来源于同一祖先。

对 apAPH 的基本酶学性质和其序列进行分析，我们发现 apAPH 还具有许多酯酶、脂肪酶家族的特征。比如其催化三联体的排列顺序（Ser445…

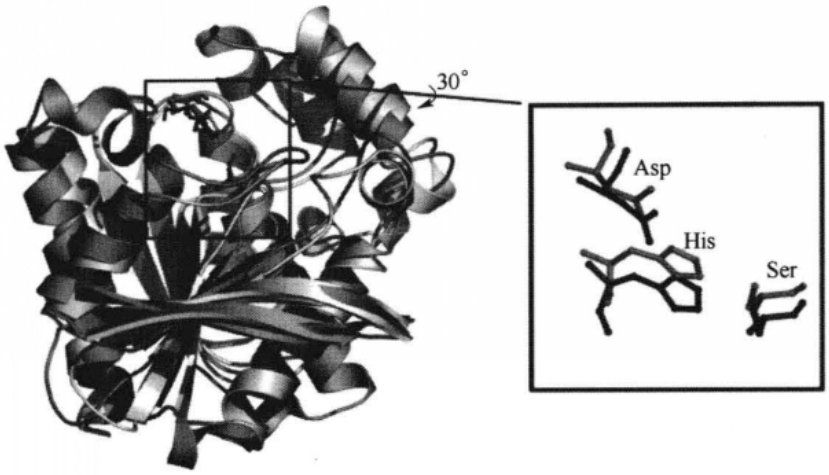



图 8.2　来自 POP 家族 apAPH（青色）和来自 HSL 家族的 AFEST（灰色）催化结构域和催化三联体的结构重叠（见彩图）

Asp524…His556）与脂肪酶、酯酶（Ser144…Asp203…His257）相同，却与经典的丝氨酸蛋白酶家族成员胰凝乳蛋白酶（His57…Asp102…Ser195）、枯草杆菌蛋白酶（Asp32…His64…Ser221）不同；在其催化活性中心 Ser 附近具有脂肪酶家族的典型基序（Gly-X-Ser-X-Gly）；更为重要的是，apAPH 表现了较高的酯酶活性，其绝对催化活力比其“天然活性”酰基肽水解酶活力还要高 8 倍左右。这些现象显示，apAPH 与脂肪酶/酯酶超家族之间具有结构、功能上的联系，以天然的 apAPH 作为进化起点，通过蛋白质工程的手段提高其酯酶活力，也许是获得具有良好性质的新型酯酶的一条崭新思路。对突变体的性质进行全面的分析，不仅可以帮助理解 apAPH 双重酶活力的结构基础，还有可能为阐明酯酶、脂肪酶家族与丝氨酸蛋白酶家族在分子进化中的联系提供新的启示。

我们对 HSL 家族与 POP 家族活性中心的残基组成进行了详细比较，发现临近催化三联体 Asp524 的 526 位点在两个家族中均高度保守，但氨基酸性质却表现出明显的差异（表 8.1）。在 POP 家族中，该位点为带电的精氨酸；但在 HSL 家族中主要为疏水残基。Arg526 位点在 POP 家族中的作用很早就受到人们的关注，来源于猪脑的 POP 晶体结构显示，Arg526 参与了与寡肽底物的结合，构成了酶与底物作用的 S2 位点（图 8.3）。在脂肪酶家族中，该位点的作用尚没有详细的研究，但这个氨基酸位于酶与底物酰基结合的通道中，也可能对底物的结合具有重要作用。由于酰基肽水解酶 apAPH 与酯酶 AFEST 的催化结构域极为相似，而在活性中心的 526 位点上却存在极性的明显差异，促使我们推测 526 位点可能对酶的底物特异性有重要的影响。因此，我们选定 apAPH 的 526 位点进行饱和突变，探讨其对于酶催化和底物选择性的作用。

**表 8.1　乙酰氨基乙酰肽酶（APH），脯氨酸寡肽酶（POP）和激素敏感脂肪酶（HSL）Arg526 位附近的序列比对**

| 酶的来源 | | 残基号① | | | | |
|---|---|---|---|---|---|---|
| | | 524 | 525 | 526 | 527 | 528 |
| APH | *Aeropyrum pernix* K1 | Asp | Ser | Arg | Thr | Phe |
| | *Caenorhabditis elegans* | Asp | Leu | Arg | Val | Val |
| | rat liver | Asp | Ala | Arg | Val | Phe |
| | human erythrocytes | Asp | Arg | Arg | Val | Phe |
| | *Bacillus subtilis* | Asp | Asp | Arg | Cys | Phe |
| | *Fugu rubribes* | Asp | Lys | Arg | Val | Phe |
| POP | *Pyrococcus horikoshii* | Asp | Asp | Arg | Val | His |
| | porcine | Asp | Asp | Arg | Val | Val |
| | *Thermobifida fusca* | Asp | Thr | Arg | Val | Asp |
| | *Gloeobacter violaceus* | Asp | Pro | Arg | Val | Asp |
| HSL | *Archaeoglobus fulgidus* | Asp | Phe | Leu | Arg | Asp |
| | brefeldin A | Asp | Phe | Leu | Arg | Asp |
| | *Alicyclobacillus acidocaldarius* | Asp | Phe | Leu | Arg | Asp |
| | Human | Asp | Phe | Met | Leu | Asp |
| | *Escherichia coli* | Asp | Phe | Leu | Leu | Asp |
| | Moraxella TA144 | Asp | Ile | Leu | Arg | Asp |

① 对应于 *Aeropyrum pernix* K1 乙酰氨基乙酰肽酶的残基号。

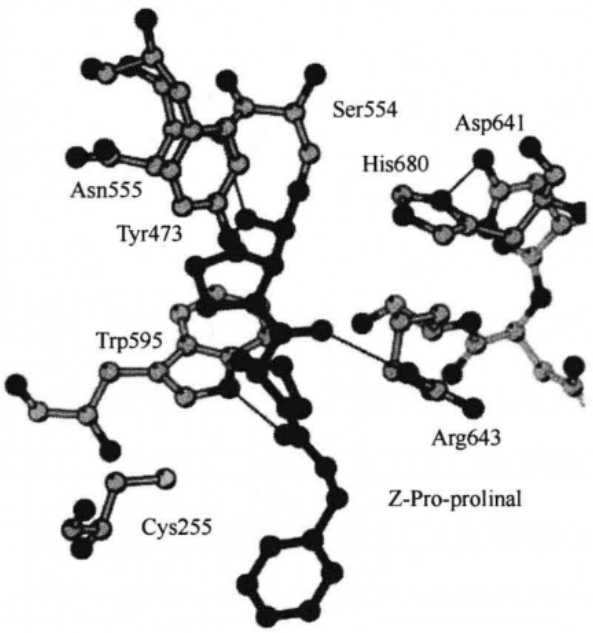



图 8.3　POP 酶的活性中心 Arg643（在 apAPH 酶中对应 Arg526）与肽底物形成一对氢键

### 8.4.3.2　Arg526 饱和突变库的建立及筛选

利用带有简并密码子的引物，通过全质粒 PCR 一次将 20 种可能的天然

氨基酸全部引入野生型 pET15b-apAPH 基因，对 Arg526 位点进行饱和突变。将 PCR 产物经过 Dpn Ⅰ 酶切消化后转化进入大肠杆菌 BL21（DE3）中构建了 Arg526 的饱和突变库。接下来利用实验室中已经建立好的、针对嗜热酯酶的高通量筛选方法，对饱和突变库的酯酶、肽酶活力进行筛选。为保证编码 20 种氨基酸的 32 种密码子在库中被筛选到的概率大于 99%，我们挑选了大约 300 个克隆用来筛选。结果发现饱和突变库中的突变体对两种底物表现出完全不同的性质（图 8.4）。大多数突变体的酯酶活力都获得了提高，其中约有一半的突变体活力比野生型 apAPH 提高 2 倍以上，约 12%的突变体活力表现为野生型活力的 5 倍以上。酯酶活力提高最明显的突变体活力约为野生型的 7 倍。令人吃惊的是，绝大多数突变体对肽酶底物 Ac-Leu-*p*NA 的活力都下降了，仅有约 20%的突变体表现为野生型 apAPH 50%以上的活力。

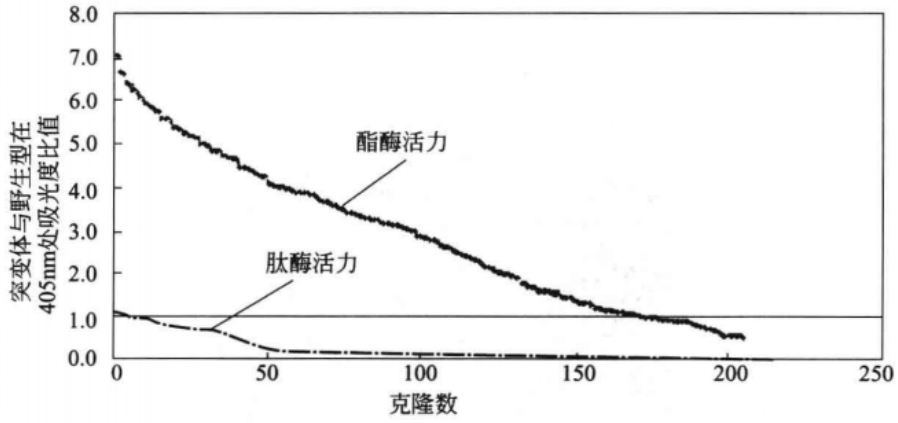



图 8.4　apAPH 饱和突变库酯酶和肽酶活力的筛选

对活性变化最明显的突变体进行测序，发现肽酶活力最低的突变体为 R526E，而酯酶活力最高的突变体由 R526I、R526L、R526V 等突变组成，证明了 526 位点在催化两种底物时起到了不同的作用。对 526 位点的任何突变都会降低 apAPH 的肽酶活力，说明经过自然选择的优化，Arg 已经是该位点的最佳选择。而酯酶活力表现出了对疏水残基很强的偏好性，说明在 526 位点疏水残基比 Arg 更适合水解酯酶底物。

#### 8.4.3.3　突变体的纯化与性质鉴定

为了详细研究 526 位点不同侧链对催化的影响，进一步构建了比 Arg 侧链短的 R526K 和完全去掉其侧链的 R526A 两个突变体。为了进一步确定突变体活力变化的详细情况，对阳性突变体进行了高效表达及蛋白质纯化，经过镍柱亲和色谱和分子筛两步纯化，野生型和突变体 apAPH 的纯度都达到了 95%以上（图 8.5、图 8.6），可以进行下一步的酶学性质表征。

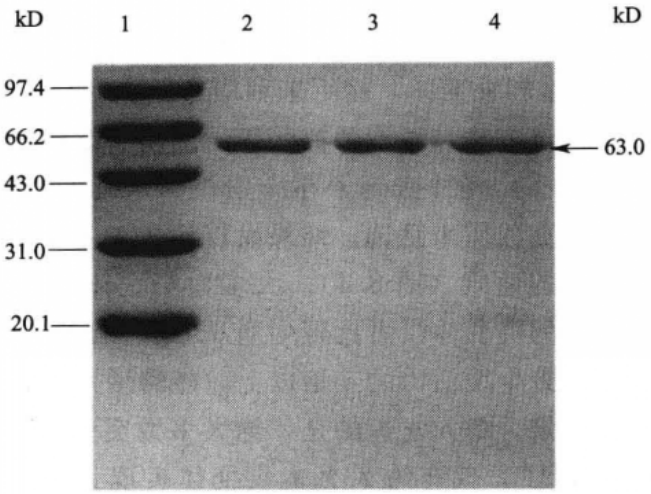



图 8.5　apAPH 纯化突变体的 SDS-PAGE 分析

1—分子量标准；2—突变体 R526I；3—突变体 R526V；4—突变体 R526L

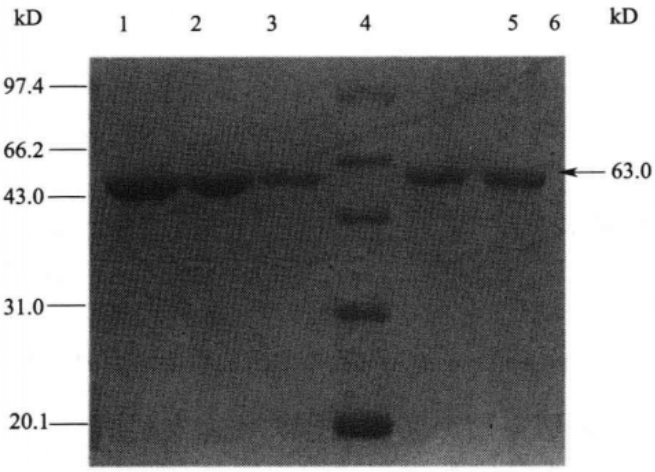



图 8.6　apAPH 纯化的野生型和突变体的 SDS-PAGE 分析

1—突变体 R526A；2—突变体 R526E；3—突变体 R526K；

4—分子量标准；5—野生型 apAPH；6—突变体 R526V

动力学测定表明，作为与野生型侧链差别最大的突变体，R526E 导致水解肽酶底物 Ac-Leu-*p*NA 时催化效率下降了 2 个数量级（表 8.2）。同时，526 位点疏水的突变 I、L、V 对酯酶底物的催化效率分别上升了 3.4 倍、4.9 倍和 5.6 倍。R526A 与 R526K 虽然侧链不同，但对两种底物的催化效率差别却不明显。

总体说来，526 位点的突变对于不同底物产生了不同的影响，在水解肽酶底物时突变体的 $K_m$ 都有明显的上升；水解酯酶底物时，尽管 R526E 的 $K_m$ 也出现了明显的增高，但催化效率的变化主要是由 $k_{cat}$ 的增高引起

的，进一步说明了 526 位点在催化两种不同类的底物时发挥了不同的作用。对于肽酶底物来说，$K_m$ 增高幅度最大的是带电荷的突变 R526K 和 R526E，而疏水侧链的突变 R526A、R526I、R526L 和 R526V 则变化较小，并且长侧链的 I、L、V 的 $K_m$ 比短侧链的 A 要低 2 倍左右。而在水解酯酶底物时，疏水并且侧链较长的突变 R526I、R526L 和 R526V 的 $k_{cat}$ 上升最为明显，清楚地显示了在水解酯酶底物时 526 位点对疏水侧链的偏嗜性。

**表 8.2　apAPH 野生型及其突变体的动力学常数**

| 酶 | 肽酶(Ac-Leu-*p*NA) | | | 酯酶(*p*NPC8) | | |
|---|---|---|---|---|---|---|
| | $k_{cat}$ /$s^{-1}$ | $K_m$ /(mmol/L) | $k_{cat}/K_m$ /[L/(mmol·s)] | $k_{cat}$ /$s^{-1}$ | $K_m$ /(mol/L) | $k_{cat}/K_m$ /[L/(mmol·s)] |
| WT | 9.3±0.5 | 0.4±0.0 | 22.3±3.8 | 6.6±0.4 | 43.3±2.8 | 152.4±17.1 |
| R526A | 12.6±0.4 | 4.3±0.3 | 2.9±0.1 | 9.6±0.2 | 41.6±2.6 | 230.7±7.1 |
| R526E | 0.5±0.0 | 5.8±0.6 | 0.1±0.0 | 8.1±0.4 | 114.6±7.7 | 70.7±9.0 |
| R526K | 26.9±1.3 | 10.5±1.2 | 2.6±0.3 | 9.4±0.6 | 38.2±2.8 | 246.1±31.5 |
| R526I | 9.2±0.5 | 2.1±0.1 | 4.4±0.6 | 20.8±1.5 | 40.1±3.6 | 518.7±5.25 |
| R526L | 9.6±0.4 | 1.3±0.0 | 7.4±0.8 | 28.5±1.1 | 38.2±2.9 | 746.1±98.5 |
| R526V | 9.7±0.3 | 1.7±0.0 | 5.7±0.6 | 30.9±0.9 | 35.7±1.6 | 865.5±14.6 |

由于 526 位点的突变同时造成酯酶活力的上升和肽酶活力的下降，所以突变体对于两种底物的选择性发生了极大的变化（图 8.7）。以野生型 apAPH 水解两种底物活力的比值为 1，突变体这个比值均大幅度提高。其中变化最大的 R526E 对酯酶底物的选择性为野生型的 129 倍，含有疏水侧链的 R526A、R526I、R526L 和 R526V 分别提高了 11.85 倍、17.38 倍、11.88 倍、22.83 倍，显示了 526 位点对于底物识别的重要作用。同时测定了突变体对不同碳链长度的酯酶底物的选择性，与野生型相比并没有变化（图 8.8），说明 526 位点仅决定肽酯两类底物的选择性，而不能影响对酯酶底物大小的选择性。

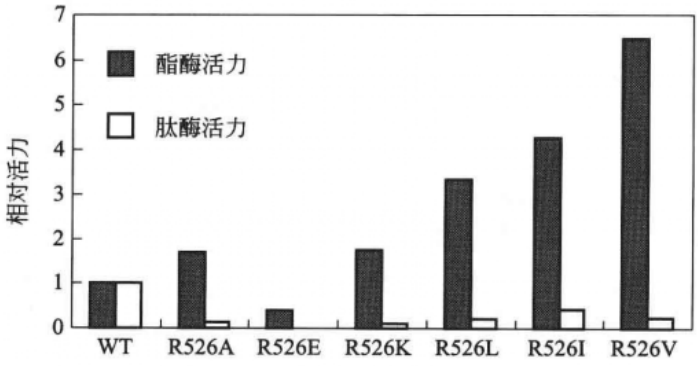



图 8.7　R526 位点突变对两种底物活力的影响

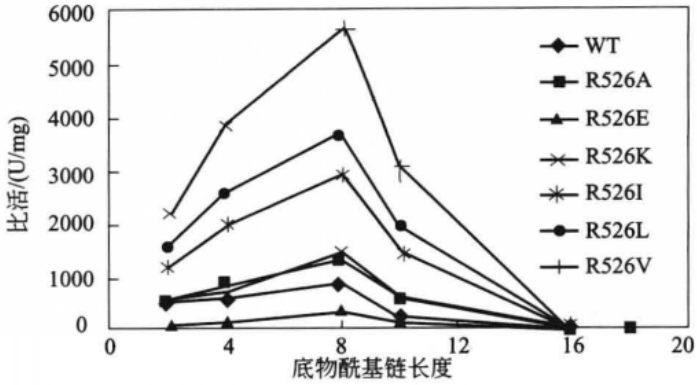



图 8.8　野生型 apAPH 和其突变体对不同链长对硝基苯酚酯的比活

#### 8.4.3.4　计算机建模

蛋白质分子的动态构象是酶发挥功能的决定性因素，利用分子动力学模拟对蛋白质在溶液中的动态行为进行分析，有助于理解突变体功能变化的机理。通过分析已知的脯氨酸寡肽酶与二肽酶Ⅳ的晶体结构，我们可以获得 apAPH 底物结合位点的足够信息，并以此为基础应用 Insight Ⅱ 软件进行分子对接，得到 apAPH 与肽酶底物 Ac-Leu-*p*NA 复合物的构象。ApAPH 肽酶底物结合口袋的 S1 亚部位临近活性中心，主要由 Met477、Phe485、Phe488、Ile489、Leu492、Trp474、Tyr446 和 Val471 等疏水残基构成，以疏水作用容纳底物 P1 位氨基酸，S2-P2 相互作用主要是 Arg526 正电荷胍基侧链与底物 P2 酰基氧通过氢键实现［图 8.9(a)］。而 apAPH 结合酯酶底物 *p*NPC8 的位点主要是由 Phe371、Phe485 和 Phe488 的侧链与 Gly368-Pro370 的主链构成，形成了明显的疏水隧道［图 8.9(b)］。

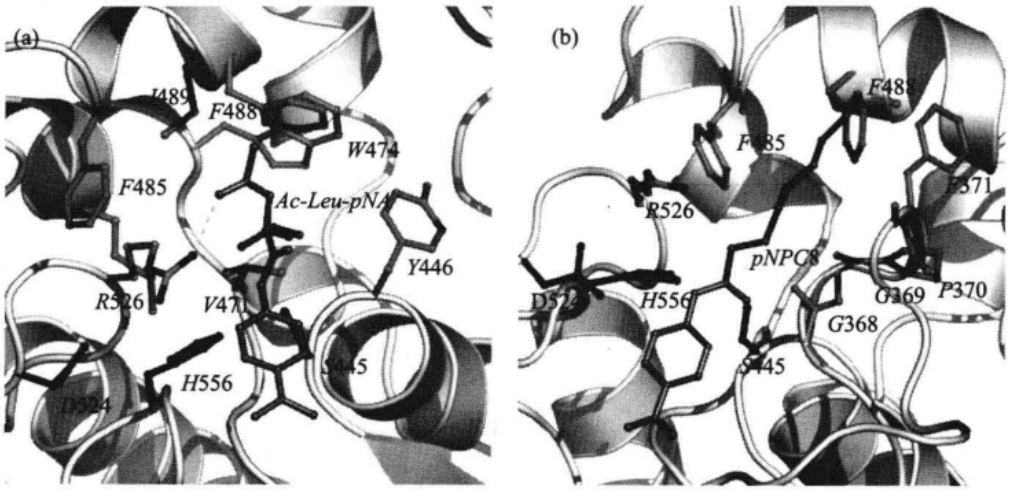



图 8.9　apAPH 的肽结合位点（a）和酯结合位点（b）

使用 PROCHEK 软件绘制对接模型的拉氏构象图，证明了酶-底物复合

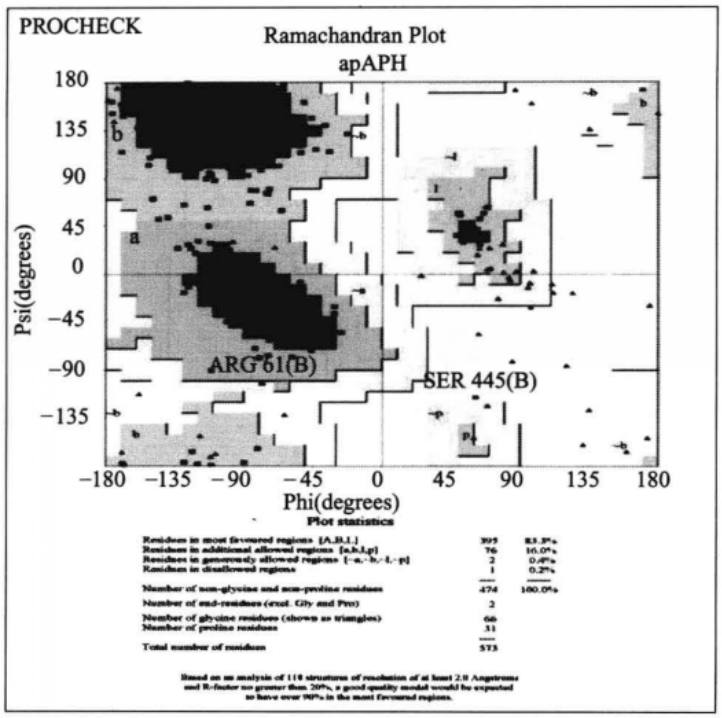



图 8.10　apAPH 的 PROCHECK 拉氏构象图（见彩图）

红色区域代表核心允许区，黄色指一般允许区，亮黄色指最大允许区，白色指不合理区

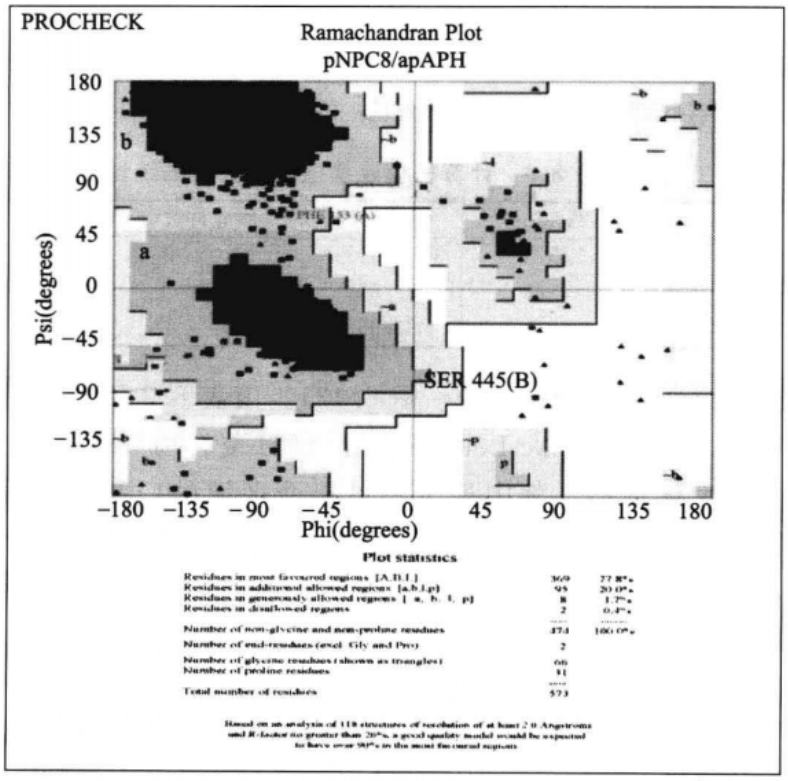



图 8.11　复合物 *p*NPC8/apAPH 的 PROCHECK 拉氏构象图（见彩图）

红色区域代表核心允许区，黄色指一般允许区，亮黄色指最大允许区，白色指不合理区

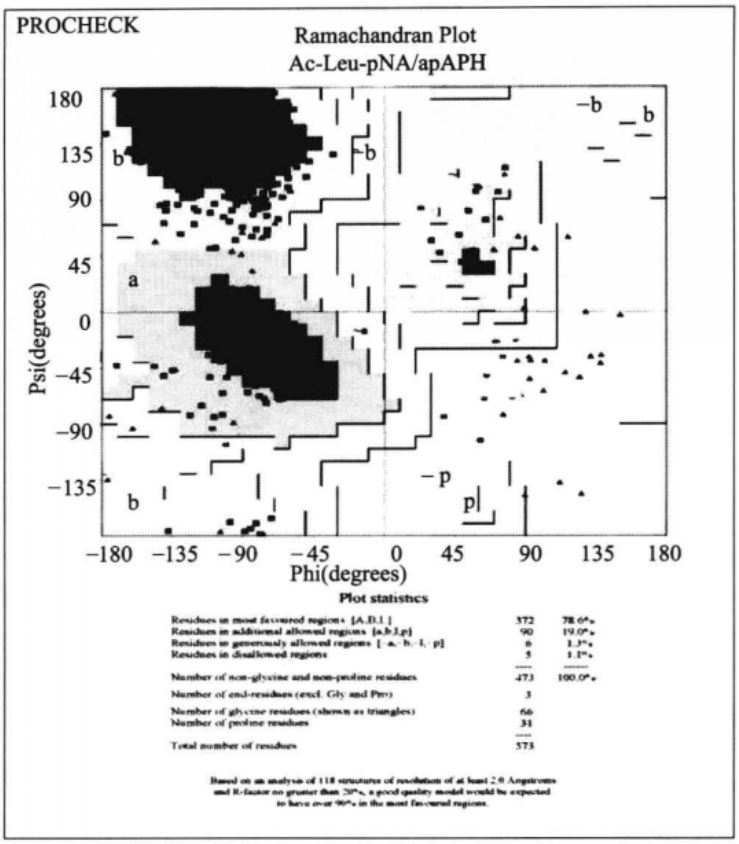



图 8.12　Ac-Leu-*p*NA/apAPH 复合物的 PROCHECK 拉氏构象图（见彩图）

红色区域代表核心允许区，黄色指一般允许区，亮黄色指最大允许区，白色指不合理区

**表 8.3　计算机模建模型的拉氏数据统计**

| 结　构 | 残 基 数 | 最合理区 | 一般允许区 | 最大允许区 | 不合理区 |
|---|---|---|---|---|---|
| apAPH | 573 | 78.6% | 19.0% | 1.3% | 1.1% |
| *p*NPC8/apAPH | 573 | 77.8% | 20.0% | 1.7% | 0.4% |
| Ac-Leu-*p*NA /apAPH | 573 | 83.3% | 16.0% | 0.4% | 0.2% |

物模拟结构的合理性（图 8.10～图 8.12）。由表 8.3 可见，野生型 apAPH 酶与 *p*NPC8/apAPH、Ac-Leu-*p*NA/apAPH 复合物三者的肽链残基二面角合理性分布大体近似，其中处于不可接受区域的残基比率分别为 1.1%、0.4%和 0.2%。与野生型相比，虽然在两种酶-底物复合物的模型内添加了小分子，但是经过能量优化后的 *p*NPC8/apAPH、Ac-Leu-*p*NA /apAPH 模型的肽链构象还要优于不带底物的空酶分子晶体结构。在酶-底物复合物的拉氏构象图中，参与底物结合的残基基本分布在最优构象区，表明构建的模型中小分子对接与蛋白质结合位点的界面构象合理。可以在此结构基础上进行突变体的建模，并进行更深入的结构分析与分子动力学模拟计算。

**8.4.3.5**　分子动力学模拟

我们选择了活力变化明显的突变体，应用分子动力学模拟对它们与酯

酶底物 *p*NPC8 和肽酶底物 Ac-Leu-*p*NA 的作用模式进行了详细研究，以期揭示活力变化的原因。*p*NPC8 复合物分子动力学模拟的结果如图 8.13 所示，野生型 Arg526 在一级序列上与活性中心的 Asp524 只间隔一个残基，其正电侧链指向活性中心 His556 上方，虽然在 *p*NPC8/apAPH 模型中其侧链并未与底物分子发生直接的接触，但是其正电胍基侧链靠近底物结合部位，有可能对底物的正确定向产生影响。在分子动力学模拟过程中，Arg526 的精氨酸侧链移动很小（RMSD 约等于 0.5Å），与底物分子的距离也保持相对稳定［图 8.13(a)］。

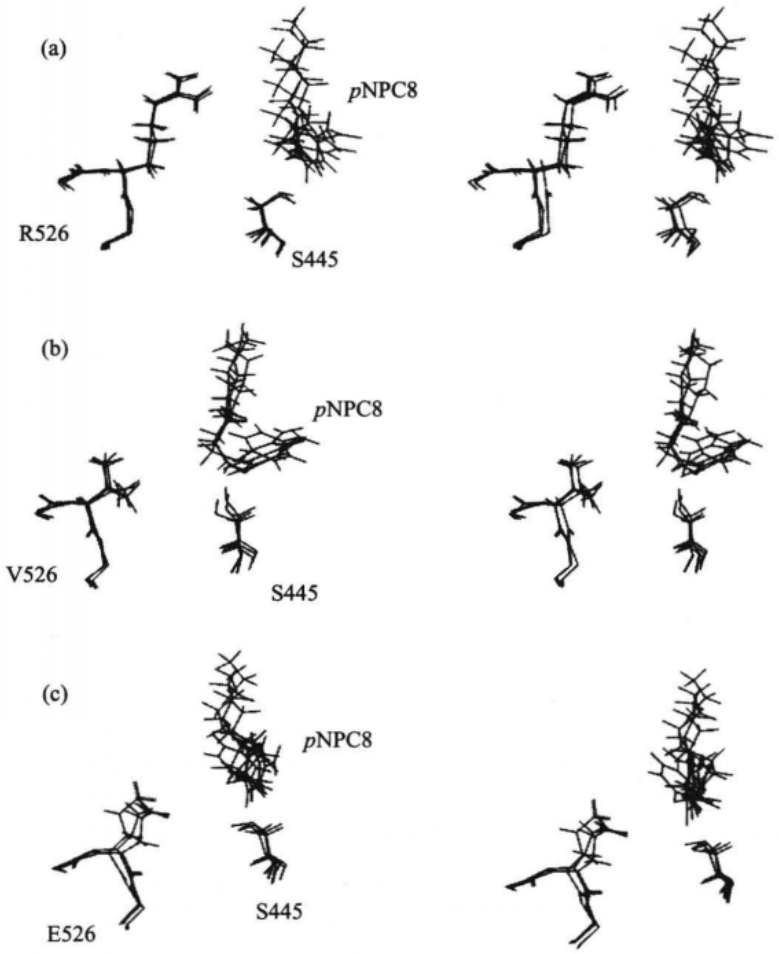



图 8.13　复合物模型 *p*NPC8/apAPH（a），*p*NPC8/R526E（b）和 *p*NPC8/R526V（c）100 皮秒分子动力学模拟构象重叠

突变体 R526V 对 *p*NPC8 的催化效率比野生型提高了约 6 倍。在 100ps 的分子动力学模拟过程中，突变的 Val 侧链表现出非常大的柔性（RMSD 约为 2.6Å，而其他的氨基酸此值均小于 1Å），由于底物的酰基侧链受到疏水的 Val 侧链牵引，导致底物分子与野生型模型相比，分子扭曲程度明显加大［图 8.13(b)］。这种作用有可能对小分子底物

的定向产生影响，使得酶与底物的相对位置和尺寸更有利于反应进行，从而使 R526V 的催化效率明显提高。

突变体 R526E 是在我们的研究中唯一表现出酯酶活力催化效率下降的突变体，酶动力学数据显示，酶和底物结合能力的下降（$K_m$ 升高）是导致其活力下降的主要原因。在 R526E 与酯酶底物的复合物模型中可见，与野生型相比，其侧链沿 $C_\alpha$-$C_\beta$ 和 $C_\beta$-$C_\gamma$ 发生较大扭转，新的侧链位置指向酯底物结合隧道，位于底物下方［图 8.13(c)］，并且在分子动力学过程中观察到其负电末端与疏水的底物酰基侧链，底物结合位点中的 Phe485 也有直接相互作用，这两种作用可能就是酶-底物结合力明显削弱，酶活力大幅度下降的主要原因。

在野生型 apAPH 与肽酶底物 Ac-Leu-*p*NA 的分子动力学模拟过程中［图 8.14(a)］，Arg526 的侧链胍基始终保持位置稳定，胍基一侧的 $N_{\eta 1}$ 与底物 P2 位的酰基氧形成氢键，另一侧 $N_{\eta 2}$ 与推进器结构域上的 Glu88 形成离子键，这个氢键网络保证了底物分子与酶形成较稳

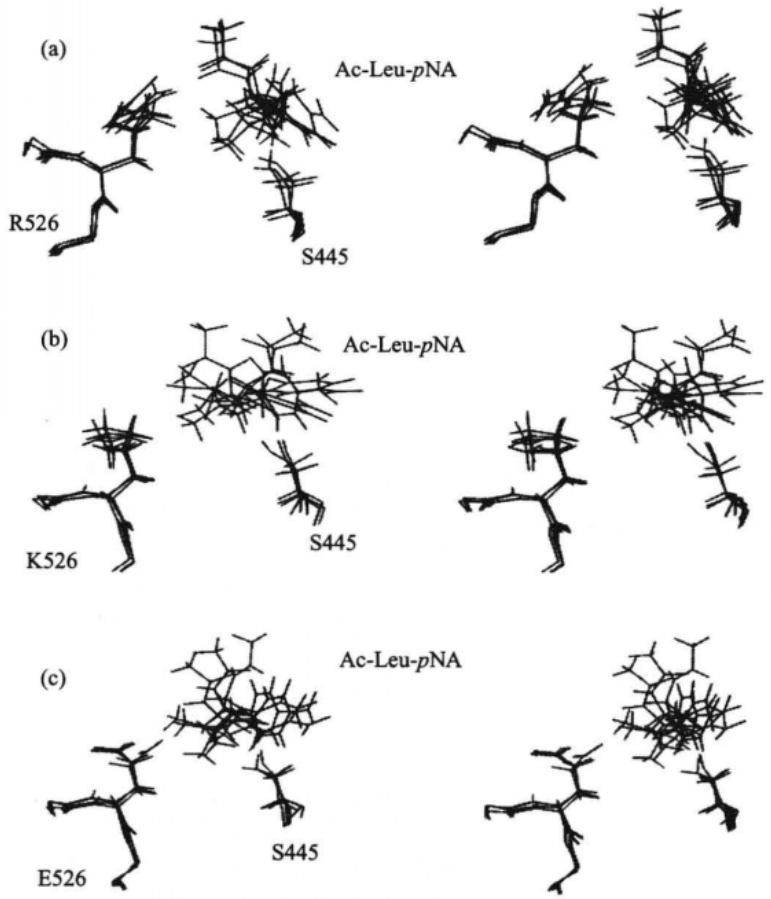



图 8.14　复合物模型 Ac-Leu-*p*NA/apAPH（a），Ac-Leu-*p*NA/R526K（b）和 Ac-Leu-*p*NA/R526E（c）100ps 分子动力学模拟构象重叠

定的复合物。在整个动力学模拟过程中底物分子的位置偏移很小（RMSD 约为 1Å），而在其他突变体的分子动力学模拟中此值均大于 1.5Å。

突变体 R526E 对肽酶底物的催化效率明显下降，其 $K_m$ 值较野生型提高了约 15 倍，$k_{cat}$降低了约 20 倍。R526E 突变破坏了野生型 S2-P2 相互作用中 Arg526 与底物分子的氢键。除此之外，在其分子动力学模拟中 R526E 的羧基侧链表现出较大的分子柔性，与底物 S1 结合位点中的 Phe485 发生了直接相互作用，降低了酶与底物结合的主要作用力，导致 $K_m$ 值的升高［图 8.14(b)］；R526E $k_{cat}$值的降低可能是由于 526 突变为 Glu 后影响了临近的催化残基 Asp524 侧链的位置，使催化三联体的正常运转受到影响而导致的。

作为与精氨酸具有类似长度和电荷性质的赖氨酸，R526K 突变同样导致酶对肽底物 $K_m$ 值的明显提高。在分子动力学模拟的过程中，Lys 侧链明显受到了 Glu88 的吸引［图 8.14(c)］，导致在整个分子动力学模拟中，始终不能与底物的酰基氧形成氢键，致使酶和底物的 S2-P2 相互作用被破坏，导致其在 $K_m$ 值上基本与侧链很小的 R526A 十分接近。这表明尽管在 20 种氨基酸里 Lys 与 Arg 最为相似，但是仍然不能替代其在此发挥作用。

### 8.4.4 讨论

Arg526 位于 apAPH 的活性中心，并且在整个脯氨酸寡肽酶（POP）家族中高度保守，说明它具有重要的作用。在猪 POP 与抑制剂的共结晶结构中发现，Arg526 对应的残基 Arg643 参与了底物的结合。通过本章描述的实验，我们确认了 Arg526 位点在 apAPH 与肽酶底物结合中具有重要的作用。实验结果表明，在 apAPH 催化肽酶底物水解时 Arg526 位点主要参与了酶与底物 S2-P2 相互作用。这种作用主要是通过 Arg 的侧链胍基集团与底物 P2 位酰基氧形成氢键而实现的。将 Arg 突变为任意一个氨基酸都会引起催化效率的下降，最佳的残基（Arg）与最差的残基（Glu）相比，催化效率相差 100 倍以上，证明 Arg 在 526 位点的保守性是由于维持功能的需要。突变体 R526K 与野生型相比，仅仅将侧链变小却使 R526K 的 $K_m$ 变为所有突变体中最大的，说明 Arg 在 526 位点的功能经过自然进化的微调已经达到最优化，任何突变都会使酶活力产生巨大的降低。类似地，R526A 与 R526I、R526L、R526V 同属不带电的疏水突变，但是 R526A 的 $K_m$ 值却高于其他侧链较大的突变 2 倍以上，表明在此位置新引入的疏水侧链，虽然稳定底物结合作用不能像 Arg 一样强，但是与完全消除侧链影响的 R526A 相比，仍较有利于底物的结合。

我们发现526位点在酯酶水解时也起到了重要的作用，带有疏水侧链的突变体水解酯酶底物的活力均明显提高，而带电荷的R526K、R526E及野生型apAPH（Arg526）与带有疏水侧链的突变体相比则活力明显降低。这个结果在分子对接中也得到了确认，在*p*NPC8/apAPH的复合物模型中，Arg526位于酯酶底物的疏水通道中，其正电侧链必定会影响疏水的酯酶底物进入活性中心，而将其突变为疏水的残基有利于酶与底物的结合。

在本实验的分子动力学模拟中发现，Arg526参与到一个由Arg105、Glu88与其共同构成的离子网络中，锚定了Arg526的侧链位置。使得在整个分子动力学模拟过程中，Arg526在两种酶底物复合物模型的侧链位置一直稳定。这个离子网络位于催化结构域与β-推进器结构域的界面上，起到了连接两个结构域的作用，在整个POP家族中保守程度很高。已有实验证明，在猪POP中这种相互作用对于肽底物进入活性中心具有重要作用，起到了控制底物进入通道的开关作用。

脯氨酸寡肽酶家族作为丝氨酸蛋白酶中新近确立的一员，自发现以来一直受到人们的关注。尽管早在1992年，Polgar就提出POP家族与酯酶家族在进化上更为相近，但是关于它们之间进化关系的线索一直缺少直接的证据。我们通过CE数据库的三级结构比较发现：apAPH催化结构域与HSL类酯酶AFEST结构相似性非常高。一般来源于一个共同祖先的后代，在序列相似性低于25%时，仍能在结构上保持主链结构偏差小于2Å。apAPH催化结构域与AFEST之间序列相似性仅为18%，可主链偏差仅为2.3Å。表明两者很有可能来源于同一祖先，并且两者都对长链酯底物表现出高水解活力。在HSL类酯酶中，Arg526的对应位点强烈偏嗜疏水氨基酸（表8.1）。结合在本试验的结果，526位点的疏水残基会显著提高apAPH对酯酶底物的催化活力，我们推测POP家族与HSL具有共同的祖先。而Arg526可能是蛋白酶和脂肪酶趋异进化中的一个进化标记位点。

### 8.4.5 小结

在本章工作中，我们发现来源于*Archeoglobus fulgidus*的激素敏感脂肪酶AFEST与apAPH的结构相似性非常高，有可能来源于同一祖先。通过对在POP家族中高度保守的Arg526位点进行饱和突变，我们证明了该位点对apAPH肽酶/酯酶活力的选择性起到了重要作用。在对其饱和突变库进行筛选的过程中，我们观察到几乎所有R526X突变体肽酶活力都出现了下降，而大部分的突变体酯酶活力升高。通过酶动力学分析与分子动力学模拟，证明了Arg526位点对肽类及酯类底物具有不同的作用方式：它参与了

酶与肽类底物的 S2-P2 相互作用，有利于肽酶底物的结合；但其带电侧链影响了疏水的酯酶底物进入活性中心，将其突变为疏水残基有利于酯酶底物的水解。本章工作提供了单点突变导致酶功能转换的稀有例证，揭示了脯氨酸寡肽酶家族与脂肪酶家族在分子进化上的相关性：在 APH 的进化历程中单一位点的突变即可使其从酯酶转换为肽酶，而 Arg526 很可能是其趋异进化的一个标记位点。

## 参考文献


[1] Masui A, Fujiwara N, Imanaka T. Stabilization and rational design of serine protease Aprm under highly alkaline and high temperature conditions. Appl Environ Microbiol, 1994, 60: 3579-3584.

[2] Morley K L and Kazlauskas R J. Improving enzyme properties: when are closer mutations better? Trends Biotechnol, 2005, 23: 231-237.

[3] Cho B, Park H, Seo J, Kim J, Kang T, Lee B, Kim B. Redesigning the substrate specificity of $\omega$-aminotransferase for the kinetic resolution of aliphatic chiral amines. Biotechnol Bioeng, 2008, 99: 275-284.

[4] Lee S, Chang Y, Shin D, Han J, Seo M, Fazelinia H, Maranas C D, Kim H. Designing the substrate specificity of D-hydantoinase using a rational approach. Enzyme Microb Technol, 2009, 44: 170-175.

[5] Bordes F, Cambon E, Dossat-Letisse V, Andre I, Croux C, Nicaud J M, and Marty A. Improvement of Yarrowia lipolytica Lipase Enantioselectivity by Using Mutagenesis Targeted to the Substrate Binding Site. Chem Bio Chem, 2009, 10: 1705-1713.

[6] Lafaquiere V, Barbe S, Puech-Guenot S, Guieysse D, Cortes J, Monsan P, Simeon T, Andre I, and Remaud-Simeon M. Control of Lipase Enantioselectivity by Engineering the Substrate Binding Site and Access Channel. Chem Bio Chem, 2009, 10: 2760-2771.

[7] Kapoli P, Axarli I A, Platis D, Fragoulaki M, Paine M, Hemingway J, Vontas J, Labrou N E. Engineering sensitive glutathione transferase for the detection of xenobiotics. Biosens Bioelectron, 2008, 24: 498-503.

[8] Jackson C J, Weir K, Herlt A, Khurana J, Sutherland T D, Horne I, Easton C, Russell R J, Scott C and Oakeshott J G. Structure-based rational design of a phosphotriesterase. Appl Environ Microbiol, 2009, 75: 5153-5156.

[9] Spadiut O, Leitner C, Salaheddin C, Varga B, Vertessy B G, Tan T, Divne C and Haltrich D, Improving thermostability and catalytic activity of pyranose 2-oxidase from Trametes multicolor by rational and semi-rational design. FEBS Journal, 2009, 276: 776-792.

[10] Spadiut O, Pisanelli I, Maischberger T, Peterbauer C, Gorton L, Chaiyen P, Haltrich D, Engineering of pyranose 2-oxidase: Improvement for biofuel cell and food applications through semi-rational protein design. J Biotechnol, 2009, 139: 250-257.

[11] Shao X, Gao Y, Jiang M, Li L. Deletion and site-directed mutagenesis of laccase from Shigella dysenteriae results in enhanced enzymatic activity and thermostability. Enzyme Microb Technol, 2009, 44: 274-280.

[12] Tran V, Hoffmann L, Rabiller C, Tellier C and Dion M. Rational design of a GH1 $\beta$-glycosi-

dase to prevent self-condensation during the transglycosylation reaction. Protein Eng Des Sel, 2010, 23: 43-49.

[13] Gribenko A V, Patel M M, Liu J, McCallum S A, Wang C, and Makhatadze G I, Rational stabilization of enzymes by computational redesign of surface charge-charge interactions. Proc Natl Acad Sci USA, 2009, 106: 2601-2606.

[14] Liu L, Wang B, Chen H, Wang S, Wang M, Zhang S, Song A, Shen J, Wu K, Jia X. Rational pH-engineering of the thermostable xylanase based on computational model. Process Biochem, 2009, 44: 912-915.

[15] Rhimi M, Aghajari N, Juy M, Chouayekh H, Maguin E, Haser R and Bejar S. Rational design of Bacillus stearothermophilus US100 L-arabinose isomerase: Potential applications for D-tagatose production. Biochimie, 2009, 91: 650-653.

[16] Ochoa-Leyva A, Soberón X, Sánchez F, Argüello M, Montero-Morán G and Saab-Rincón G. Protein Design through Systematic Catalytic Loop Exchange in the $(\beta/\alpha)_8$ Fold, J Mol Bio, 2009, 387: 949-964.

[17] Mileni M, Johnson D S, Wang Z, Everdeen D S, Liimatta M, Pabst B, Bhattacharya K, Nugent R A, Kamtekar S, Cravatt B F, Ahn K and Stevens R C, Structure-guided inhibitor design for human FAAH by interspecies active site conversion. Proc Natl Acad Sci USA, 2008, 105: 12820-12824.

[18] Kamondi S, Szilágyi A, Barna L, Závodszky P. engineering the thermostability of a TIM-barrel enzyme by rational family shuffling. Biochem Biophys Res Commun, 2008, 374: 725-730.

[19] Boersma Y L, Pijning T, Bosma M S, van der Sloot A M, Godinho L F, Droge M J, Winter R T, van Pouderoyen G, Dijkstra B W, and Quax W J. Loop Grafting of Bacillus subtilis Lipase A: Inversion of Enantioselectivity. Chemistry & Biology, 2008, 15: 782-789.

[20] Höst G E, Jonsson B. Converting human carbonic anhydrase Ⅱ into a benzoate ester hydrolase through rational redesign. Biochim Biophys Acta, 2008, 1784: 811-815.

[21] Jochens H, Stiba K, Savile C, Fujii R, Yu J, Gerassenkov T, Kazlauskas R J, and Bornscheuer U T, Converting an Esterase into an Epoxide Hydrolase. Angew Chem Int Ed, 2009, 48: 3532-3535.

[22] Murphy P M, Bolduc J M, Gallaher J L, Stoddard B L, and Baker D. Alteration of enzyme specificity by computational loop remodeling and design. Proc Natl Acad Sci USA, 2009, 106: 9215-9220.

[23] Yeung N, Lin Y, Gao Y, Zhao X, Russell B S, Lei L, Miner K D, Robinson H and Lu Y. Rational design of a structural and functional nitric oxide reductase. Nature, 2009, 462: 1079-1084.

[24] Caldwell R C, Joyee G F. Randomization of genes by PCR mutagenesis. PCR Methods Appl, 1992, 2: 28-33.

[25] Shao Z, Zhao H, Giver L. Random priming in vitro recombination: an effective tool for directed evolution. Nucl Acids Res, 1998, 26: 681-685.

[26] Sieber V, Martinez C A, Arnold F H. Libraries of hybrid proteins from distantly related sequence. Nat Biotechnol, 2001, 19: 456-460.

[27] Zhao H, Giver L, Shao Z. Molecular evolution by staggered extension process (StEP) in vitro recombination. Nat Biotechnol, 1998, 16: 258-261.

[28] Coco W M. RACHITT: Gene family shuffling by Random Chimeragenesis on Transient Tem-

plates. Methods Mol Biol, 2003, 231: 111-127.

[29] Horswill A R, Naumann T A, Benkovic S J. Using incremental truncation to create libraries of hybrid enzymes. Methods Enzymol, 2004, 388: 50-60.

[30] Moore G L, Maranas C D, Gutshall K R. Modeling and optimization of DNA recombination. Computers and Chemical Engineering, 2000, 24: 693-699.

[31] Voigt C A, Mayo S L, Arnold F H, Wang Z G. Computational method to reduce the search space for directed protein evolution. Proc Natl Acad Sci USA, 2001, 3: 778-783.

[32] FernanAdez Gacio A, Uguen M, Fastrez J. Phage display as a tool for the directed evolution of enzymes. Trends Biotechnol, 2003, 21: 408-414.

[33] Schaffitzel C, Hanes J, Jermutus L. Ribosome display: an in vitro method for selection and evolution of antibodies from libraries. Journal of Immunological Methods, 1999, 231: 119-135.

[34] Lingen P. Improving the carboligase activity of benzoylformate decarboxylase from Pseudomonas putida by a combinationof directed evolution and site-directed mutagenesis. Protein Eng, 2002, 15: 585-593.

[35] Griffiths A D and Tawfik D S. Directed evolution of an extremely fast phosphotriesterase by in vitro compartmentalization. EMBO J, 2003, 22: 24-35.

[36] Stemmer W P C. DNA shuffling by random fragmentation and reassembly in vitro recombination for molecular evolution. Proc Natl Acad Sci USA, 1994, 91: 747-751.

[37] Arnold F H, Giver L, Gershenson A, Zhao H, Miyazaki K. Directed evolution of mesophilic enzymes into their thermophilic counterparts. AnnNY Acad Sci, 1999, 18: 400-403.

[38] Giver L, Gershenson A, Freskgard P O, Arnold F H. Directed evolution of a thermostable esterase. Proc Natl Acad Sci U S A, 1998, 95: 12809-12813.

[39] Komeda H. Enhancement of the thermostability and catalytic activity of D-stereospecific amino acid amidase from Ochrobactrum anthropi SV3 by directed evolution. J Mol Catal B. Enz, 2003, 21: 283-290.

[40] Ji-Hu Zhang, Glenn Dawes, and Willem P C Stemmer. Directed evolution of a fucosidase from a galactosidase by DNA shuffling and screening. Proc Natl Acad Sci U S A, 1997, 94: 4504-4509.

[41] Alexeeva M. Deracemisation of a-methylbenzylamine using an enzyme obtained by in vitro evolution. Angew Chem Int Ed Engl, 2002, 41: 3177-3180.

[42] May O. Inverting enantioselectivity by directed evolution of hydantoinase for improved production of L-methionine. Nat Biotechnol, 2000, 18: 317-320.

[43] Jaeger K E, Reetz M T. Directed evolution of enantioselective enzymes for organic chemistry. Curr Opin Chem Biol, 2000, 4: 68-73.

[44] Williams G J. Modifying the stereochemistry of an enzyme-catalyzed reaction by directed evolution. Proc Natl Acad Sci U S A, 2003, 100: 3143-3148.

[45] Chica R A, Doucet N, Pelletier J N. Semi-rational approaches to engineering enzyme activity: combining the benefits of directed evolution and rational design. Curr Opin Biotechnol, 2005, 16: 378-84.

[46] Santoro S W, Schultz P G. Directed evolution of the site specificity of Cre recombinase. Proc Natl Acad Sci USA, 2002, 99: 4185-4190.

[47] Rui L, Cao L, Chen W, Reardon K F, Wood T K. Active site engineering of the epoxide hydrolase from Agrobacterium radiobacter AD1 to enhance aerobic mineralization of *cis*-1,2-di-

chloroethylene in cells expressing an evolved toluene ortho-monooxygenase. J Biol Chem, 2004, 279: 46810-46817.

[48] Reetz M T. Controlling the enantioselectivity of enzymes by directed evolution: practical and theoretical ramifications. Proc Natl Acad Sci USA, 2004, 101: 5716-5722.

[49] Geddie M L, Matsumura I. Rapid evolution of β-glucuronidase specificity by saturation mutagenesis of an active site loop. J Biol Chem, 2004, 279: 26462-26468.

[50] Peimbert M, Segovia L. Evolutionary engineering of a β-lactamase activity on a D-Ala D-Ala transpeptidase fold. Protein Eng, 2003, 16: 27-35.

[51] Hayes R J, Bentzien J, Ary M L, Hwang M Y, Jacinto J M, Vielmetter J, Kundu A, Dahiyat B I. Combining computational and experimental screening for rapid optimization of protein properties. Proc Natl Acad Sci USA, 2002, 99: 15926-15931.

[52] Gao R, Feng Y, Ishikawa K, Ishida H, Ando S, Kosugi Y, Cao S, Cloning, purification and properties of a hyperthermophilic esterase from archaeon Aeropyrum pernix K1. Journal of Molecular Catalysis B: Enzymatic, 2003, 24-25: 1-8.

[53] Bartlam M, Wang G, Yang H, Gao R, Zhao X, Xie G, Cao S, Feng Y, Rao Z. Crystal structure of an acylpeptide hydrolase/esterase from Aeropyrum pernix K1. Structure, 2004, 12: 1481-1488.

[54] Shindyalov I N, Bourne P E. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng, 1998, 11: 739-747.

[55] Zheng L, Baumann U, Reymond J L. An efficient one-step site-directed and site-saturation mutagenesis protocol. Nucleic. Acids Res, 2004, 32: e115.

[56] Laskowski R A, MacArthur M W, Moss D S, Thornton J M. PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Cryst, 1993, 26: 283-291.

[57] Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins, 1991, 9: 56-68.

[58] Doolittle R F. Similar amino acid sequences: chance or common ancestry. Science, 1981, 214: 149-159.

[59] Russell R B, Barton G J. Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chain contacts secondary structure and accessibility. J Mol Biol, 1994, 244: 332-350.

[60] Grishin N V. Estimation of evolutionary distances from protein spatial structures. J Mol Evol, 1997, 45: 359-369.

[61] Michael L Sierk and William R Pearson. Sensitivity and selectivity in protein structure comparison. Protein Sci, 2004, 13: 773-785.

[62] Chothia C, and Lesk A M. The relation between the divergence of sequence and structure in proteins. EMBO J, 1986, 5: 823-826.

[63] Gao R, Feng Y, Ishikawa K, Ishida H, Ando S, and Ksougi Y. Cloning, purification and properties of a hyperthermophilic esterase from archason Aeropyrum pernix K1. J Mol Catal B Enzym, 2003, 24-25: 1-8.

[64] Nardini M, Lang D A, Liebeton K, Jaeger K E, and Dijkstra B W. Crystal structure of Pseudomonas aeruginosa lipase in the open conformation: the prototype for family I of bacterial lipases. J Biol Chem, 2000, 275: 31219-31225.

[65] Rawlings N D, Polgár L, and Barrett A J. A new family of serine-type peptidases related to

prolyl oligopeptidase. Biochem J, 1991, 279: 907-908.

[66] Arrett A J, and Rawlings N D. Oligopeptidases, and the emergence of the prolyl oligopeptidase family. Biol Chem Hoppe Seyler, 1992, 373: 353-360.

[67] Polgár L. The prolyl oligopeptidase family. Cell Mol Life Sci, 2002, 59: 349-362.

[68] Szeltner Z, Rea D, Renner V, Juliano L, Fülöp V, and Polgár L. Electrostatic Environment at the Active Site of Prolyl Oligopeptidase Is Highly Influential during Substrate Binding. J Biol Chem, 2003, 278: 48786-48793.